IS SI

2021

18th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

ISSI2021

Virtual event

12–15 July 2021
KU Leuven, Belgium

PROCEEDINGS

ISSI

2021

18th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

ISSI2021

Virtual event

12–15 July 2021
KU Leuven, Belgium

PROCEEDINGS

18th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

ISSI2021

12–15 July 2021
KU Leuven, Belgium

PROCEEDINGS

Editors
Wolfgang Glänzel, Sarah Heeffer, Pei-Shan Chi, Ronald Rousseau

ISBN 9789080328228 – ISSN 2175-1935
July 2021 Printed in Belgium

Sponsors and Funding

Diamond sponsor – Elsevier
Gold sponsor – Digital Science
Gold sponsor – RISIS
Silver sponsor – JDIS
Silver sponsor – MIT Press
Silver sponsor – Clarivate ISI
Bronze sponsor – Springer Nature
Bronze sponsor – FRONTIERS

Research Foundation – Flanders (FWO)

Organisers

Koenraad Debackere, KU Leuven (Belgium), *Chair*
Wolfgang Glänzel, KU Leuven (Belgium)
Bart Thijs, KU Leuven (Belgium)
Tim Engels, UAntwerp (Belgium)

Programme Committee: Chairs

Ronald Rousseau, KU Leuven (Belgium), *Chair*
Lin Zhang, Wuhan University (China)/KU Leuven (Belgium)
Henk F. Moed, Sapienza University of Rome (Italy)

Program Committee: Members

Kevin Boyack, SciTech Strategies Inc. (USA), *Chair Workshops and tutorials*
Claudia Gonzalez Brambila, ITAM (Mexico)
Julie Callaert, KU Leuven (Belgium)
Pei-Shan Chi, KU Leuven (Belgium), *Poster Session*
Cinzia Daraio, Sapienza University of Rome (Italy)
Jacqueline Leta, UFRJ (Brazil), *Poster Session*
Raf Guns, UAntwerp (Belgium)
Stefanie Haustein & Juan Pablo Alperin, Ottowa (Canada)
Sarah Heeffer, KU Leuven (Belgium)
Omwoyo Bosire Onyancha, University of South Africa (South Africa)
Walter Ysebaert, VUB (Belgium)
Lin Zhang, Wuhan University (China)/KU Leuven (Belgium)

Doctoral Forum Committee

Jens Peter Andersen, Aarhus University (Denmark), *Chair*
Guillaume Cabanac, University of Toulouse (France)
Vincent Larivière, Université de Montréal (Canada)
Andrea Scharnhorst, KNAW (Netherlands)
Cassidy Sugimoto, Indiana University (USA)
Alesia Zuccala, University of Copenhagen (Denmark)

Eugene Garfield Doctoral Dissertation Award 2021 Committee

Birger Larsen, University of Aalborg (Denmark), *Chair*
Elías Sanz Casado, Carlos III University of Madrid (Spain)
Gemma Derrick, Lancaster University (UK)
Nees Jan van Eck, Leiden University (Netherlands)

Committee for the ISSI Student Travel Award

Kevin Boyack, SciTech Strategies Inc. (USA), *Chair*
Aparna Basu, South Asian University (India)
Cinzia Daraio, Sapienza University of Rome (Italy)

Scientific Committee

Abramo Giovanni
Adams Jonathan
Adie Euan
Aguillo F. Isidro
Ahlgren Per
Ajiferuke Isola
Akbaritabar Aliakbar
Aman Valeria
Amez Lucy
Andersen Jens Peter
Archambault Eric
Arvanitis Rigas
Astrom Fredrik
Ausloos Marcel
Azagra-Caro Joaquín M.
Badin Luiza
Baiget Tomas
Ball Rafael
Ballester-Gonzalez Omar
Bartol Tomaz
Bauer Guntram
Benner Mats
Bhattacharya Sujit
Bianchi Federico
Bollen Johan
Bonaccorsi Andrea
Borner Katy
Bornmann Lutz
Boshoff Nelius
Bouabid Hamid
Bowman Timothy
Buehrer Susanne
Cabezas-Clavijo Álvaro
Calero-Medina Clara
Campbell David
Cañibano Carolina
Cao Cong
Carayol Nicolas
Casnici Niccolò
Catalano Giuseppppe
Caviggioli Federico
Chang Ching-Chun
Chaurasia Neeraj
Chavarro Diego
Checchi Daniele
Chen Chaomei
Chen Dar-Zen
Chen Yue
Chincilla-Rodriguez Zaida
Ciarli Tommaso
Colavizza Giovanni
Confraria Hugo
Costas Rodrigo
Côté Grégoire
Criado Regino
D’Angelo Ciriaco Andrea
Dastidar Prabir G.
De Bellis Nicola
De Bordes Carole
De Moya Felix
Dehdari-rad Tahereh
Deinzer Gernot
Díaz-Faes Adrián A.
Ding Ying
Donner Paul
Fatchur Rochim Adian
Ferreira-Goncalves Márcia
Franssen Thomas
Frietsch Rainer
Furner Jonathan
Gallart Jordi Molas
Garcia Romero Antonio
Gauch Stephan
Getz Daphne
Ghiasi Gita
Glänzel Wolfgang
Gläser Jochen
Gok Abdullah
Gorraiz Juan
Gorry Philippe
Gregori Martina
Guevara Miguel
Gumpenberger Christian
Gzoyan Edita
Halevi Gali
Hammarfelt Björn
Haunschild Robin
Hinze Sybille
Holmberg Kim
Hörlesberger Marianne
Hornbostel Stefan
Hu Xiaojun
Huang Mu-Hsuan
Ingwersen Peter
Ivancheva Ludmila
Jana Siladitya
Jeng Wei
Jimenez-Contreras Evaristo
Jonkers Koen
Jovanovic Milos
Kajikawa Yuya
Katz Sylvan
Kelchtermans Stijn
Klavans Dick
Konkiel Stacy
Kousha Kayvan
Krauskopf Manuel
Kulczycki Emanuel
Kumaravel J P S
Lamers Wout
Laudel Grit
Laurens Patricia
Lawson Cornelia
Lemke Steffen
Leng Rhodri
Lepori Benedetto
Li Jiang
Licea Judith
Liu Yuxian
Lo Szu-Chia

López-Illescas Carmen
Luwel Marc
Maisano Domenico
Markusova Valentina
Martín-Martín Alberto
Maynard Diana
Mayr Philipp
Mêgnigbêto Eustache
Meho Lokman I.
Melkers Julia
Meoli Michele
Milojevic Stasa
Mongeon Philippe
Morillo Fernanda
Moskaleva Olga
Mouton Maria
Mugabushaka Alexis Michel
Mugnaini Rogerio
Must Ülle
Mutz Ruediger
Nane Tina
Nederhof Anton
Nieminen Pentti
Noyons Ed
Ochsner Michael
Olmeda-Gómez Carlos
Orozco Luis Antonio
Ortega José Luis
Pandiella Andres
Pellegrino Gabriele
Perianes-Rodríguez Antonio
Peritz Cheila, Bluma
Peset Fernanda
Peter Viola
Peters Isabella
Pinheiro Diogo
Pislyakov Vladimir
Pontille David
Porter Alan L.
Pouris Anastassios
Prathap Gangan
Prozesky Heidi
Qiu Junping
Raffo Julio D.
Rafols Ismael
Rao Ravichandra
Reale Emanuela
Rigby John
Rigolin Camila Carneiro Dias
Robinson Garcia Nicolas
Roche Ivana
Romanello Matteo
Ronda-Pupo Guillermo
Rons Nadine
Rotolo Daniele
Rousseau Ronald
Rousseau Sandra
Ruggiero John
Ruocco Giancarlo
Sachwald Frederique
Sahoo Bibhuti
Sainte-Marie Maxime
Salinas Daniel
Sangam Shivappa Lingappa
Sargsyan Shushanik
Scherngell Thomas
Schiebel Edgar
Schmidt Marion
Schmoch Ulrich
Schneider Jesper
Schubert András
Schubert Torben
Seeber Marci
Shapira Philip
Shen Zhesi
Shu Fei
Sivertsen Gunnar
Small Henry
Smit Nynke-Jo
Soós Sándor
Sorz Johannes
Srivastava Divya
Sterzi Valerio
Sun Yuan
Tang Li
Tesch Jakob
Thelwall Mike
Traag Vincent
Tunger Dirk
Van den Besselaar Peter
Van der Meulen Barend
Van Leeuwen Thed
Van Looy Bart
Van Raan Anthony
Vancauwenbergh Sadia
Vargas-Quesada Benjamin
Velden Theresa
Vinkler Peter
Visser Martijn
Wagner Caroline
Waltman Ludo
Wang Lili
Wang Qi
Wang Xianwen
Welch Eric
Winnink Jos
Wolfram Dietmar
Woolley Richard
Wouters Paul
Wroblewski Angela
Wu Jinshan
Wu Yishan
Yan Erjia
Yang Liying
Ye Fred Y.
Yegros Alfredo
Youtie Jan
Zahedi Zohreh
Zhao Dangzhi
Zhao Yajuan
Zhou Ping
Zirulia Lorenzo

PREFACE

The 18th International Society of Scientometrics and Informetrics Conference is held on 12–15 July 2021 and organized by KU Leuven in close collaboration with the university of Antwerp under the auspices of ISSI the International Society for Scientometrics and Informetrics.

Due to the constraints imposed by the COVID-19 pandemic situation and the resulting risks for planning and organising a full-fledged conference with physical presence of attendees, the organisers decided to go virtual this time. This decision warranted a smooth organisation without unpredictable events and unnecessary modifications and adjustments during the preparation process. However, organising a virtual event of this size has posed a true challenge for the organisers. We were very pleased that despite the various different time-zones in the world, we could organise part of the conference, notably all keynotes, invited talks and plenary sessions as well as the workshops and tutorials and, last not least, all ceremonies as live events. The special tracks and the usual "parallel sessions" had, however, to be organised using pre-recorded on-line video streams. Making a virtue of necessity, we can provide all talks and the discussion for three months after the conference and thus make the conference a hopefully sustainable event.

The goal of ISSI2021 is to provide an international open forum for scholars and practitioners in the domain of informetrics, bibliometrics, scientometrics, webmetrics and altmetrics to discuss new research directions, advanced methods and theories, and to highlight the best research in this area. In order to achieve this goal, we asked researchers worldwide to submit original research manuscripts, particularly full papers, research-in-progress papers or posters, to propose and organise tutorials and workshops, with a special emphasis on the future of this area and on its interdisciplinary links with other fields. We succeeded in attracting a sufficient number of contributions to organise two special tracks, one on bibliometric approaches to measure and evaluate interdisciplinary research (IDR) organised by KU Leuven and one on the bibliometrics of social sciences and humanities (SSH) organised by our partner at University Antwerp. University Antwerp also took the opportunity to introduce their project of publishing on a new handbook on "Research Assessment in the Social Sciences".

We would like to thank the three keynote speakers, who have accepted our invitation, namely Katy Börner from Indiana University in Bloomington (USA), Albert-László Barabási from Northeastern University in Boston (USA), and David Sweeney, Executive Chair of Research England (UK).

Despite the pandemic-induced constraints, we could accept about 230 contributions, roughly two thirds of which are full and research-in-progress papers and nearly one third is present in pre-recorded poster sessions.

The assignment of contributions to session topics was facilitated by the organisation as a virtual event since time constraints did not influence the schedule. We have chosen the following major themes.

- Research evaluation & bibliometrics in support of science policy
- Effects of research funding
- Patent analysis
- Individual-level bibliometrics
- Collaboration and mobility
- Domain studies and regional issues
- Open Science, Open Access and editorial impact
- Gender and equality studies
- Webmetrics, altmetrics and media impact of research
- Advanced methods in citation analysis
- Data sources and data processing
- Document and journal analysis
- Network analysis

Furthermore, six workshops deal with the tracing of epistemic change, with creation and application of models, cited reference analysis, national and institutional research assessment, and collaborative archive and data research environment. This is completed by two tutorials. The organisers look forward to the future publication of the outcomes of these workshops.

All accepted presentations and posters are incorporated in the conference proceedings. The first part comprises invited and full papers as well as research-in-progress papers, while the second part is devoted to posters.

At this place, we would like to express our thanks to all participants and contributors for their understanding for the constrains due to the special situation and their active support. Our thanks are due to the ISSI board for its trust and support. We also thank the Leuven Congress Office – PC for their enormous effort and dedicated work and MEEPLE for providing the online platform. In particular, we thank Liesbeth Michiels, Marie-Laure Bettens and Ann Moerenhout.

On behalf of the organisers and the programme committee of ISSI2021
Wolfgang Glänzel

INDEX OF PAPERS
(FULL PAPERS AND RESEARCH IN PROGRESS)

INDEX OF POSTERS

Do hijacked journals attract dishonest authors?

Anna Abalkina[1]

[1] *abalkina@gmail.com*
Free University of Berlin (Germany)

Abstract

This research in progress studies academic misconduct as it related to hijacked journals. There is a common belief in the literature that naïve authors submit papers to fake journals. This study demonstrates that there is a group of dishonest authors who are attracted by the fast publication with no peer review process that is offered by hijacked journals. The results show that the average share of plagiarism in an examined sample of 500 papers was 21.0%. Plagiarism was not detected only in 28.2% of the papers. These results raise concerns not only about cyber-criminals but also about authors who exploit hijacked journals to improve their publication records.

Introduction

Hijacked journals create a challenge for academic publishing. Hijacked journals represent a type of cyber-crime. Hijacked journals mimic legitimate journals (Lukić 2014; Bohannon 2015; Dadkhah 2015; Jalalian & Dadkhah 2015; Asadi et al. 2017; Shahri et al. 2018) in order to cheat potential clients. These cloned journals exploit the titles and ISSNs of original journals and fraudulently collect fees from authors.

There is a common belief that the potential clients of such journal are naïve authors who are not able to distinguish between honest and fraudulent journals (Watson 2015; Dadkhah& Borchardt 2016). However, is this true? In other words, are naïve authors the only group of authors who submit their articles to hijacked journals? This question has not yet been investigated in the literature. There is an alternative hypothesis that some authors choose to exploit hijacked journals and submit their articles in order to increase their publication records. There is evidence of some plagiarism cases in hijacked journals and of the recycling of already published texts to replenish the archives of fraudulent journals (Abalkina 2021). Dadkhah et al. (2016) detected some cases of the circulation of texts between predatory and hijacked journals. Abalkina (2020) demonstrated cases of translation plagiarism in papers submitted to the hijacked *Journal of Talent Development and Excellence*.

In this study, I argue that dishonest authors constitute another group of clients who submit their articles to hijacked journals. Dishonest authors are attracted by fraudulent hijacked journals that offer a fast publication process with no peer review. To test this hypothesis, the texts of papers published in hijacked journals were checked for text similarities. Plagiarism is considered to be one of the most serious forms of academic misconduct (Resnik et al. 2015). Authors who violate academic ethics can be considered dishonest. Plagiarism detection in articles submitted to hijacked journals can shed light on the behaviour of the authors who submit to such journals.

This study is important for several reasons. First, there is a rising concern about the proliferation of hijacked journals and fraudulent publishers (Dadkhah & Borchardt 2016; Memon 2019). Second, recent evidence suggests that cyber-criminals compromise information from the webpages of peer-review journals and their content in international citation databases (Al-Amr 2020). Such behaviour constitutes a significant challenge for the academic community. Third, a considerable amount of literature has developed around the topic of academic misconduct and plagiarism in scientific papers (Fanelli 2009; Pupovac & Fanelli 2015). However, these studies do not take into account possible violations of academic ethics in hijacked journals. Fourth, research has shown that the high level of competition in academia, as well as the 'publish or perish' strategy, can increase the number of publications in predatory journals (Kurt 2018). Publishing in hijacked journals can be another possible way for academics to improve their publication records and cheat their universities.

Selection of hijacked journals

The hijacked journals were selected from several sources. First, I used the lists of hijacked journals found on the online websites https://beallslist.net/hijacked-journals/ and https://predatoryjournals.com/hijacked/. Second, I used the lists of hijacked journals found in Jalalian & Dadkhah (2015) and Abalkina (2021).

Third, I searched the SCImago website, which contains the profiles of academic journals and allows comments by users. Several hijacked journals were selected according to the comments of users regarding hijacked journals.

All the websites of identified hijacked journals were checked for their availability because hijacked journals do not work for a long time, and their domains usually expire after several years. I have made a list of 85 cloned websites that were available as of March 2021.

Selection of articles

The selection of articles to check for plagiarism is challenging. There is evidence that the archives of cloned journals are replenished by texts from hijacked and predatory journals (Abalkina 2021). To select the papers that were submitted to the hijacked journals but were not added by hijacking themselves, I selected only the last three issues of hijacked journals. This method increased the probability of selecting papers that were submitted by authors. I selected each tenth paper of each issue to check for plagiarism. I randomly chose the first paper (from one to ten) and then downloaded each tenth paper. If the total number of papers was less than ten, I chose each fifth paper. In the case of the *Journal of Talent Development and Excellence*, each twentieth paper was downloaded due to the extreme number of papers in each issue of this fraudulent journal.

Plagiarism detection

I used Urkund (Ouriginal) to detect plagiarism in the papers published in the hijacked journals. The study by Foltýnek et al. (2020) confirmed the efficiency of Urkund in detecting text similarities. Since the software detects all possible text similarities (Weber-Wulff 2019), a manual check is needed to eliminate the published paper itself, the references, citations, and standard phrases, for example, the description of statistical models, as well as false positive results. A manual check is also needed to detect the issue date of the text donors to take into consideration reverse plagiarism. After the manual check of text similarities, the final percentage of non-originality was calculated. As of March 2021, I have analysed a sample of 500 papers (approximately 60%).

Results

This study in progress found that 1,150 authors that contributed to 500 papers in the hijacked journals were predominantly from developing or emerging countries, especially from India (see Figure 1). However, this authorship by scientists from India (71,1%) is explained by the fact that a large number of hijacked journals are aimed at authors from India. Each tenth author of the sample papers originated from Indonesia. Authors from Russia and Kazakhstan shared 3rd and 4th place. There was also evidence of plagiarism in papers authored by post-Soviet scholars and published in hijacked journals (Abalkina 2020). Interestingly, the coauthors of most papers originated from the same country, and international collaboration can be considered an exception.

The analysis of text similarities in the hijacked journals showed that the level of plagiarism in the sample was high. The current average level of nonoriginal text is 21.0%; this number will be adjusted at the end of the study.

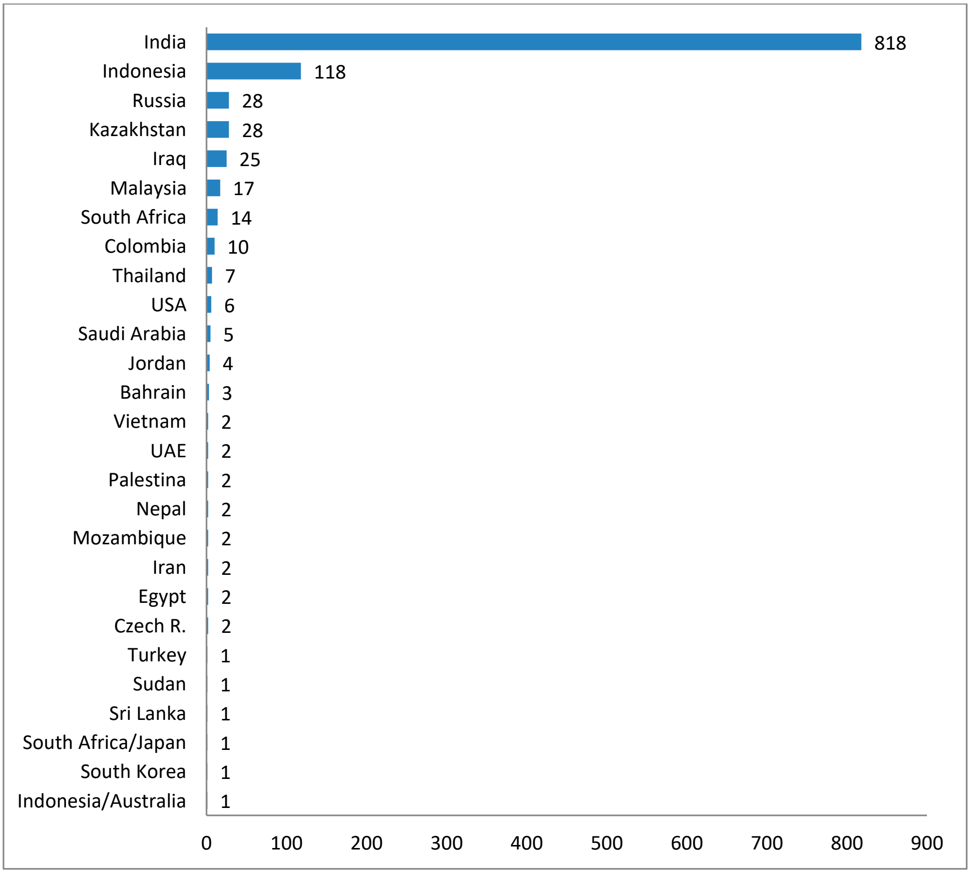


Figure 1Distribution of affiliation country of authors

Several types of plagiarism were found in the articles under investigation:

- Blatant plagiarism. Approximately 18.2% of papers had text similarities at the level of 50% or more;
- Plagiarism with rewriting. This is the case when the text of other authors is adopted, and synonyms are used;
- Plagiarism from papers with a different set of coauthors. This happens when text similarities are detected in papers with an overlapping set of coauthors;
- Self-plagiarism.
- Data fabrication. Several cases of data fabrication were also detected. In these cases, authors have changed the data in the plagiarized texts to keep them updated.

The text donors represent a variety of sources: master's degree theses, PhD theses, working papers, papers from both legitimate journals and predatory and hijacked journals, books, newspapers, and Internet pages.

Limitations

Both self-plagiarism and a low level of other-author plagiarism can be explained not only by the dishonesty of the authors but also by a low level of awareness of academic ethics principles and citation rules by such authors. This aspect warrants further investigation by interviews with the authors.

Conclusions

The following conclusions can be drawn from this study of plagiarism in hijacked journals. First, this study confirms the hypothesis that the authors are not as naïve as it is commonly believed in the literature. The absence of strict requirements for publication and plagiarism checks allows authors to publish texts with a significant amount of plagiarism. Thus, these findings of the current study suggest that fraudulent journals attract dishonest authors.

Second, the hypothesis of the presence of honest authors who have been deceived by hijacked journals is also not excluded. However, the share of articles without plagiarism is currently only 28.2%; this figure will be adjusted after the end of the study.

Acknowledgements

I thank Valeria Morozhnikova for her assistance with data collection. I gratefully acknowledge financial support from Ludwig Maximilian University of Munich where this study was initiated.

References

Abalkina, A. (2021). Detecting a network of hijacked journals by its archive. Retrieved 14.03.2021 from URL: https://arxiv.org/abs/2101.01224.

Abalkina, A. (2020). The case of the stolen journal. Retraction Watch, July 7. Retrieved 13.02.2021 from URL: https://retractionwatch.com/2020/07/07/the-case-of-the-stolen-journal/

Al-Amr M. (2020). How did content from a hijacked journal end up in one of the world's most-used databases? Retraction Watch. September 1. Retrieved 13.02.2021 from URL: https://retractionwatch.com/2020/09/01/how-did-content-from-a-hijacked-journal-end-up-in-one-of-the-worlds-most-used-databases/

Asadi, A., Rahbar, N., Asadi, M. et al. (2017). Online-based approaches to identify real journals and publishers from hijacked ones. *Science and Engineering Ethics*, 23, 305–308.

Bohannon, J. (2015). How to hijack a journal. *Science*, 350(6263), 903–905.

Dadkhah, M. (2015). New types of fraud in the academic world by cyber criminals. *Journal of Advanced Nursing*, 72, 2951-2953.

Dadkhah, M. & Borchardt, G. (2016). Hijacked journals: An emerging challenge for scholarly publishing. *Aesthetic Surgery Journal*, 36(6) 739–741.

Dadkhah, M., Maliszewski, T. & Teixeira da Silva, J.A. (2016). Hijacked journals, hijacked web-sites, journal phishing, misleading metrics, and predatory publishing: actual and potential threats to academic integrity and publishing ethics. *Forensic Science*, *Medicine*, and *Pathology*, 12, 353–362.

Fanelli, D. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLoS ONE*, 4(5), e5738.

Foltýnek, T., Dlabolová, D., Anohina-Naumeca, A. et al. (2020). Testing of support tools for plagiarism detection. *International Journal* of *Educational Technology* in *Higher Education*, 17, 46.

Jalalian M. & Dadkhah M. (2015). The full story of 90 hijacked journals from August 2011 to June 2015. *Geographica Pannonica*, 19(2), 73–87.

Kurt, S. (2018). Why do authors publish in predatory journals? *Learned Publishing*, 31, 141–147.

Lukić, T., Blešić, I., Basarin, B., Ivanović, B. L., Milošević, D. & Sakulski, D. (2014). Predatory and fake scientific journals/publishers: A global outbreak with rising trend: A review. *Geographica Pannonica*, 18(3), 69–81.

Memon, A. (2019). Hijacked journals: A challenge unaddressed to the developing world. *Journal of the Pakistan Medical Association,* 69(10),1413-1415.

Pupovac, V., Fanelli, D. (2015). Scientists Admitting to Plagiarism: A Meta-analysis of Surveys. *Science and Engineering Ethics*, 21, 1331–1352.

Resnik, D.B., Rasmussen, L.M. & Kissling, G.E. *(*2015*).* An international study of research misconduct policies*. Accountability in Research,* 22*(*5), 249–266*.*

Shahri, et al. (2018). Detecting hijacked journals by using classification algorithms. *Science and Engineering Ethics*, 24, 655–668.

Watson, R. (2015). Hijackers on the open access highway. *Nursing Open,* Nov; 2(3): 95–96.

Weber-Wulff, D. (2019). Plagiarism detectors are a crutch, and a problem. *Nature*, 567(7749), 435.

Do the propensity and drivers of academics' engagement in research collaboration with industry vary over time?

Giovanni Abramo[1], Francesca Apponi[2] and Ciriaco Andrea D'Angelo[3]

[1] *giovanni.abramo@iasi.cnr.it*
Laboratory for Studies in Research Evaluation, Institute for System Analysis and Computer Science (IASI-CNR), National Research Council of Italy (Italy)

[2] *francesca.apponi@alumni.uniroma2.eu*
Department of Engineering and Management, University of Rome "Tor Vergata" (Italy)

[3] *dangelo@dii.uniroma2.it*
Department of Engineering and Management, University of Rome "Tor Vergata" (Italy)
&
Laboratory for Studies in Research Evaluation, Institute for System Analysis and Computer Science (IASI-CNR), National Research Council of Italy (Italy)

Abstract

This study is about public-private research collaboration. In particular, we want to measure how the propensity of academics to collaborate with their colleagues from private firms varies over time and whether the typical profile of such academics change. Furthermore, we investigate the change of the weights of main drivers underlying the academics' propensity to collaborate with industry. In order to achieve such goals, we apply an inferential model on a dataset of professors working in Italian universities in two subsequent periods, 2010-2013 and 2014-2017. Results can be useful for supporting the definition of policies aimed at fostering public-private research collaborations, and should be taken into account when assessing their effectiveness afterwards.

Introduction

The ability of industry to exploit the results of academic research is a distinctive competence of advanced economies. Policies aimed at developing such ability are among the priorities of an increasing number of governments (Fan, Yang, & Chen, 2015; Shane, 2004). Public-private research collaboration is one of the main modes to realize knowledge transfer. Understanding the motivations underlying joint cooperation is an important step towards formulating policies and initiatives aimed at increasing the frequency of collaboration.

The main objective of this work is to investigate to what extent frequency of public-private research collaboration change over time, alongside the main drivers underlying the academics' propensity to collaborate with industry. Decision makers then would be able to formulate incentive schemes based on those drivers that show not only more weight but also more stability. A critical step in the study is the identification of the main drivers that could influence the propensity of academics to engage in research collaborations with industry. Previous literature suggests that they are to be found in the individual characteristics of the academic and the environment he or she works in (Zhao, Broström, & Cai, 2020; Llopis, Sánchez-Barrioluengo, Olmos-Peñuela, & Castro-Martínez, 2018; Abramo, D'Angelo, & Murgia, 2013), although personal characteristics seem to be more important than those of the environment (D'Este & Patel, 2007).

Most works on the topic are based on surveys, which poses limits on the scale of observations (Zhao, Broström, & Cai, 2020; Weerasinghe & Dedunu, 2020; Llopis, Sánchez-Barrioluengo, Olmos-Peñuela, & Castro-Martínez, 2018; Thune, Reymert, Gulbrandsen, & Aamodt, 2016). To overcome these limits, we adopt instead a bibliometric approach, analyzing the co-

authorships of publications. Co-authorship in fact should be the outcome of a real research collaboration, though exceptions cannot be excluded (Katz & Martin, 1997).
We aim to answer the following research questions:

- Does the intensity of public-private research collaboration varies over time?
- Does the typical profile of academics collaborating with the private sector change over time?
- Do the main individual and contextual drivers of the academics' propensity for research collaboration with the private sector change over time?

To address these questions, we apply an inferential model on a dataset of professors working in Italian universities in two subsequent periods, 2010-2013 and 2014-2017. It must be noted that we investigate the scientific output of research collaboration, therefore output at time *t* refers to a joint research project conducted at time t-τ, where τ represents the time lapse needed for the new knowledge to be encoded in written form and having it published in a scientific journal. This is it be kept in mind in general when interpreting results of time-series or before-after analyses, and especially when setting up frameworks to assess the effectiveness of policies aimed at fostering public-private research collaborations.
In the next section we review the literature on the factors determining private-public research collaboration. In Section three we present the methodology and data. In Section four we show the results of the analysis. Section five concludes the work.

Literature review

The underlying reasons for academics to engage in research collaborations with industry differ from those of the industrial partner. The latter mainly aims to access knowledge to be further developed for commercial exploitation (Bekkers & Bodas Freitas, 2008; Perkmann et al., 2013). The former is mostly attracted from direct economic and financial benefits that the industrial partner can offer (Garcia, Araújo, Mascarini, Santos, & Costa, 2020) and, to some extent, also by a noble purpose: the so called "mission" motivation, i.e. advancing the societal role of universities (Iorio, Labory, & Rentocchini, 2016). The intensity of public-private research collaboration must be sensitive then to the changing financial needs of academics, to the incentive schemes adopted by national and regional governments and by the research organizations they work at, and to the effectiveness of industry-liaison offices which act as catalysts for collaboration.
From the academic standpoint, the propensity to collaborate with industry varies along the career cycle, although the impact of age seems non-linear (Weerasinghe & Dedunu, 2020). In fact, collaboration propensity seems stronger at the early career stage and weaker in the later stages (Bozeman & Gaughan, 2011; Ubfal & Maffioli, 2011). In fact, younger academics are in more need of financial resources, and to show their ability to activate and manage collaborations, which is positively evaluated for career progress (Bayer & Smart, 1991; Traoré & Landry, 1997).
Gender also plays an important role. Because of gender homophily women show greater difficulty in developing their social capital, including collaboration networks (Boschini & Siogren, 2007). As compared to males, female academics engage less in research collaboration with industry (Tartari & Salter, 2015), mainly because they are less interested (Calvo, Fernández-López, & Rodeiro-Pazos, 2019).
There occurs a positive correlation between research performance and collaboration (Mansfield, 1995; Mansfield & Lee, 1996; He, Geng, & Campbell-Hunt, 2009; Lee & Bozeman, 2005; Schartinger, Schibany, & Gassler, 2001), as companies are likely to search for top players to engage in their research projects (Balconi, & Laboranti, 2006).

The tendency to diversify one's own research activities may underly curiosity for novelty and unexplored, and reveal more appropriate for engaging in cross-sector research (Abramo, D'Angelo, & Di Costa, 2019).

Environmental factors must have an influence as well on the academic's propensity for joint research with industry. The relative importance that the university poses on technology transfer creates a culture and motivational stimuli more or less harbinger of research collaboration with industry (Di Gregorio & Shane, 2003; Giuri, Munari, Scandura, & Toschi, 2019). Also, the extent of financial resources made available for research affect the propensity to search for private sources (Giuri, Munari, Scandura, & Toschi, 2019).

The attitude and inclination of academic colleagues towards cross-sector research projects, will have an influence of one's own behavior, through a mechanism of social comparison (Tartari, Perkmann, & Salter, 2014). The size of the university plays a role in determining the types of collaboration by the research staff. If a critical research mass is not there, academics would likely tend to engage in extra-muros collaborations (Schartinger, Schibany, & Gassler, 2001).

The demand by industry for collaboration, and therefore the concentration of private R&D in the territory of the university, affects the possibilities for cross-sector collaboration (Berbegal-Mirabent, Sánchez García, & Ribeiro-Soriano, 2015).

Finally, the research production function varies across scientific disciplines (Abramo, D'Angelo, & Murgia, 2013a), therefore discipline effects on the intensity of university-industry collaborations are expected.

Data and methods

The field of observation consists of professors of Italian universities, conducting research in the so-called hard sciences. We exclude the social sciences and arts & humanities because for these the coverage of bibliographic repertories is still insufficient for reliable representation of research output. In the Italian university system all professors are classified in one and only one field (named the scientific disciplinary sector, SDS, 370 in all). Fields are grouped into disciplines (named university disciplinary areas, UDAs, 14 in all). We observe research production in 201 SDSs falling in 10 UDAs, where publications in international journals serve as a reliable proxy for overall research output.

To investigate whether the collaboration rates and determinants of academic engagement in joint research with industry change over time, we conduct a longitudinal analysis dividing the observed period into two subsequent four-year subperiods, 2010-2013 and 2014-2017.

The analysis dataset consists of all assistant, associate and full professors, on staff for at least three years in 2010-2013 or 2014-2017, with at least one Web of Science (WoS) indexed publication in the relevant period. These are 31634 professors in the first period and 31392 in the second.

The bibliometric dataset is extracted from the Italian Observatory of Public Research, a database developed and maintained by the present authors, and derived under license from the Thomson Reuters WoS. Beginning from the raw data of the WoS and applying a complex algorithm to reconcile the author's affiliation and disambiguation of the true identity of the authors, each publication (article, article review and conference proceeding) is attributed to the professors that authored it (D'Angelo, Giuffrida, & Abramo, 2011).[1] Collaboration with industry is evidenced by the presence of at least one private company[2] in the address list of publications authored by the professor in the dataset.

[1] The harmonic average of precision and recall (F-measure) of authorships, as disambiguated by the algorithm, is around 97% (2% margin of error, 98% confidence interval).

[2] Identified through the manual scrutiny and unification of all bibliographic addresses with "Italy" as affiliation country.

We will first verify if the share of professors collaborating with industry in the two periods increases or not and if the "profile" of those engaging in such collaborations has changed over time. Then, we will use a logit regression in order to understand if the main drivers of academic engagement in public-private research collaboration changed in weight, in the two periods. The logit regression relies on a dummy dependent variable (Y) assuming: 1, if professor *i* co-authored at least one publication with industry; 0, otherwise. As for covariates suggested by previous literature that are likely to affect the propensity of professors to engage in research collaboration with industry, we will consider the following, grouped in two clusters.

Individual covariates

- Gender (X_1), specified by a dummy variable (1 for female; 0 for male);
- Age ($X_{2\text{-}5}$), specified with 5 classes, through 4 dummies (baseline "Less than 40");
- Academic rank ($X_{6\text{-}7}$), specified by 2 dummies (baseline "Assistant professors");
- Total publications authored by the professor in the period under observation (X_8);
- Level of specialization of the scientific activity of the professor, specified by 2 dummies:
  - ✓ "Highly diversified" (X_9), 1 if the papers falling in the prevalent subject category of the professor[3] is less than 40% of total publications; 0, otherwise;
  - ✓ "Highly specialized" (X_{10}), 1 if the papers falling in the prevalent subject category of the professor is more than 75% of total publications; 0, otherwise;[4]

Contextual covariates

- Environment-Peers behaviour (X_{11}), specified by a dummy variable (1 in case of a colleague in the same university and SDS of the professor, co-authoring publications with industry; 0, otherwise)[5];
- Institutional control (X_{12}), specified by a dummy variable (0, for public universities; 1, for private ones);
- University scope (X_{13}), specified by a dummy variable (1, for "Polytechnics" and "Special Schools for Advanced Studies, SS"; 0, otherwise);[6]
- University size in the ADU of the professor ($X_{14\text{-}15}$), specified with 3 classes through 2 dummies ("Big", for universities with a research staff in the UDA of the professor, above 80 percentile in the national ranking; "Medium", with a research staff between 50 and 80 percentile);
- University location ($X_{16\text{-}19}$), specified with 5 geographical macro-areas, by 4 dummies (baseline "Islands").
- Period (X_{20}), specified by a dummy variable (0, for 2010-2013; 1, for 2014-2017)

In order to control for the research discipline effects, we also consider other 9 dummies related to the ten UDAs under observation.

[3] For a thorough explanation of the bibliometric approach to discriminate specialized from diversified research, we refer the reader to Abramo, D'Angelo, and Di Costa (2017).

[4] The chosen thresholds (40% for X_8, and 75% for X_9) allow for equal partitions of the dataset (one third of of highly diversified professors, and one third of highly specialized ones).

[5] For this variable we do not exclude those publications where the scientist under observation is a co-author. Such publications are fruit of cross-sector and, at the same time, of "intramural" collaborations, where the peers effect is in place.

[6] The Italian Minister of University and Research (MUR) recognizes a total of 96 universities as having the authority to issue legally recognized degrees. Of these, 29 are small, private, special-focus universities, of which 13 offer only e-learning, 67 are public and generally multi-disciplinary universities. Three of them are Polytechnics and six are *Scuole Superiori* (Special Schools for Advanced Studies), devoted to highly talented students, with very small faculties and tightly limited enrolment per degree program. In the overall system, 94.9% of faculty are employed in public universities (0.5% in *Scuole Superiori*).

The logit regression was applied separately to the two period under observation. In detail, values of the variables referred to the first period are measured at 31/12/2009, while the ones of the second period are measured at 31/12/2013.

Results

In the first period, professors with at least one publication co-authored with industry represented 17.2 percent of the 2010-2013 dataset (Table 1). In the second period, their share raised to 25.5 percent. Among the 27,323 professors on staff in both periods, 10.0 percent collaborated with industry in both periods; 8.1 percent collaborated in the first period only, and 16.0 percent collaborated in the second period only. Overall, the share of professors on staff in both periods, and who collaborated with industry in the first period was 18.1 percent; in the second period, 26.0 percent.

The other two subsets considered in Table 1 (A not B, B not A), consist respectively of professors who left the academia in the second period, and who joined it in the second period. Among the former, the share of those who co-authored at least one publication with industry in 2010-2013 was 11.3 percent. Among the latter, the share almost doubled (21.6%).

We can say then, that i) the share of professors on staff in both periods, collaborating with industry increased by 8.0 percent; and ii) the new entries (in the second period) showed a propensity to collaborate with industry which almost doubled that of professors approaching retirement.[7]

Table 1. Breakdown of professors in the dataset

Set*	No.	Category	Share
A	31643	Collaborating	17.2%
B	31392	Collaborating	25.5%
A ∩ B	27323	Collaborating in both periods	10.0%
		Collaborating only in the second period	16.0%
		Collaborating only in the first period	8.1%
		Never collaborating	65.8%
A not B	4320	Collaborating with industry	11.3%
B not A	4069	Collaborating with industry	21.6%

** "A" = professors in the 2010-2013 dataset; "B" = professors in the 2014-2017 dataset*

Table 2 reports the typical profile of university professors collaborating with private companies, identified by the concentration index on each individual and contextual trait.[8] The left side refers to the profile emerging from the data of the first four-year period; the right side, to the second one.

2010-2013, the typical academic active in research collaboration with industry is male, 40-45 years old, a full professor, with highly diversified research activity. This professor operates within a group of peers who likewise collaborate with industry, and belongs to a medium-sized public university, typically a polytechnic or SSs located in northwestern Italy.

In 2014-2017, the profile is very similar. The only significant differences are in the age of the academic and size of their home university. As for the age, the prevailing trait is below 40 years old: in other words, in the second period, interaction with industry extends to and features academics younger than other aged colleagues. As for the size, the shift is from medium to big, but in both period the association is not statistically significant. Other traits remain unchanged,

[7] Almost all professors present only in the first period, were not there in the second period because of age limits. In Italy, it is quite rare that a professor quits academia for other reasons (Abramo, D'Angelo, & Rosati, 2016).

[8] The concentration index is the ratio of two ratios. Example: for the group variable "gender", the prevailing trait "male" shows a concentration index of 1.058 for 2010-2013 data, since males compose 71.66% of total researchers co-authoring publications with industry, and 67.75% of the total population, therefore 71.66/67.75=1.058.

although variation in the absolute value of the concentration index reveals a weakening/strengthening of the characteristic trait between the two periods.

Since this purely descriptive analysis does not take into account the simultaneous effects of all covariates on the independent variable, in the following we conduct an inferential analysis using a logit regression model, as illustrated in the previous section.

Table 2: Profiling of the Italian academic co-authoring publications with industry

	2010-2013			2014-2017		
Group variable	Prevailing trait	Concentr. index	Pearson chi^2	Prevailing trait	Concentr. index	Pearson chi^2
Gender	Male	1.058	46.1***	Male	1.076	120.0***
Age	40-45 years old	1.073	54.9***	Below 40 years old	1.132	132.1***
Academic rank	Full professor	1.160	78.2***	Full professor	1.150	80.9***
UDA	9 - Ind.+Infor. Engineering	1.661	1.0e+03***	9 - Ind.+Infor. Engineering	1.777	1.9e+03***
Scientific activity	Highly diversified	1.328	356.5***	Highly diversified	1.220	306.3***
Environment	Peers collaborating	1.432	1.6e+03***	Peers collaborating	1.265	1.5e+03***
University type	Public	1.005	4.6**	Public	1.012	35.6***
	Polytechnic or SS	1.619	152.3***	Polytechnic or SS	1.676	309.0***
University size	Medium	1.014	0.7	Big	1.005	1.9
University location	North-west	1.262	167.2***	North-west	1.148	149.1***

*Statistical significance: *p-value <0.10, **p-value <0.05, ***p-value <0.01*

Table 3 shows the average values of the model variables, at overall level, for the two periods under investigation. Some significant differences between the two periods, for individual covariates, deserve an underlining:

- Average age of professors increased, in particular the incidence of the elder courts (for example, the over sixties' share raised from 12.6 percent to 14.3 percent of total);
- Average number of publications per professors increased (from 14.60 to 19.37);
- The propensity to diversify research activity increased. Highly diversified academics raised from 14.3 percent to 19.1 percent. Specularly, the share of specialized researchers decreased from 31.3 percent to 25.2 percent).

We remind the reader that the two datasets partly overlap, with about 87 percent of professors belonging to both. As a consequence, variations between the two periods can be explained partly by a behavioural change of incumbent professors, and partly by different attitudes and interests of younger newly hired professors as compared to their older peers who retired. In fact, in the second period, the share of associate professors, and more noticeably of full professors, dropped in favour of that of assistant professors. In spite of that, the average age of faculty increased, in particular that of the elder courts (the share of over sixty years old professors raised from 12.6 percent to 14.3 percent). Thanks to the new entries and exits, gender unbalance improved, as the share of female professors raised from 32.3 percent to 33.9 percent. In fact, in the subset "B not A" of Table 1, the share of females is 37.6 percent, while in the subset "A not B", it is 25.3 percent.

Table 3: Average values of the regression model variables in the two periods under investigation

Variable group			2010-2013	2014-2017
Response	Y	Co-authorships with industry	0.172	0.255
Gender	X_1	Female	0.323	0.339
Age	X_2	40-45	0.200	0.190
	X_3	46-52	0.244	0.252
	X_4	53-60	0.229	0.258
	X_5	Over 60	0.126	0.143
Academic rank	X_6	Associate prof.	0.284	0.281
	X_7	Full prof.	0.247	0.217
Producivity	X_8	Tot. publications	14.60	18.37
Scientific activity	X_9	Highly diversified	0.143	0.191
	X_{10}	Highly specialized	0.313	0.252
Environment	X_{11}	Peers behaviour	0.561	0.660
University type	X_{12}	Private	0.034	0.039
	X_{13}	Polytechnic or SS	0.057	0.059
University size	X_{14}	Medium	0.329	0.321
	X_{15}	Large	0.598	0.610
University location	X_{16}	South	0.200	0.200
	X_{17}	Center	0.254	0.250
	X_{18}	Northeast	0.197	0.196
	X_{19}	Northwest	0.239	0.244

Findings of the logit regressions applied to the whole dataset to investigate the drivers of academic engagement in public-private research collaboration, are reported in Table 4.

The model estimation appears satisfactory. The mean VIF is 2.42, with maximum (7.23) for the covariate "University size – Big" (X_{15}), which excludes the presence of multicollinearity that could disturb the estimation of the coefficients.

The value of under ROC area (AUC) is 0.747, which indicates good ability of the model to correctly classify professors, discriminating the propensity to collaborate with companies[9] in function of individual traits and context.

The estimated coefficients of the regression model are expressed in terms of odds ratios: the reference value is equal to one and indicates that the independent variable considered has no effect on the dependent variable, i.e. on the probability that a professor has or has not collaborated with private companies. For values above one, the variable instead has a positive marginal effect, and vice versa.

Scrolling data in Table 4, the coefficient of X_{20}, representing the time pattern, immediately jumps to the eye: all others being equal, the propensity to collaborate in the second period increases by over 50% as compared to the first, confirming what has already been shown by descriptive statistics.

In general, all the covariates but very few have a statistically significant effect. In particular, "University type-Polytechnic or SS" (X_{13}) does not show a significant marginal effect on academics' propensity to collaborate with industry.

[9] The AUC analysis evaluates a classifier's ability to discern between true positives and false positives. In our case, the AUC value, between 0 and 1, is equivalent to the probability that the result of the logit classifier applied to a researcher randomly extracted from the group of those who collaborated with industry is higher than that obtained by applying it to a researcher randomly extracted from the group of those who did not collaborate (Bowyer, Kranenburg, & Dougherty, 2001).

Table 4: The main drivers of the propensity to collaborate with industry by Italian professors. Logit regression, dependent variable: 1 in case of publications in co-authorship with industry, 0, otherwise

			odd	Std Err.	z	P>|z|	Interaction effect
Variable group		Const.	0.041	0.003	-39.52	0.000	with X_{20} - coeff.†
Gender	X_1	Female	0.882	0.021	-5.19	0.000	-0.070
Age	X_2	40-45	0.947	0.033	-1.59	0.112	0.076
	X_3	46-52	0.784	0.029	-6.63	0.000	-0.052
	X_4	53-60	0.628	0.026	-11.37	0.000	-0.048
	X_5	Over 60	0.462	0.023	-15.27	0.000	-0.009
Academic rank	X_6	Associate prof.	1.354	0.039	10.4	0.000	-0.026
	X_7	Full prof.	1.809	0.065	16.48	0.000	-0.010
Producivity	X_8	Total publications	1.011	0.000	26.81	0.000	-0.001*
Scientific activity	X_9	Highly diversified	1.346	0.037	10.86	0.000	-0.031
	X_{10}	Highly specialized	0.620	0.017	-17.65	0.000	0.075
Environment	X_{11}	Peers collaborating	3.124	0.084	42.53	0.000	-0.172***
University type	X_{12}	Private	0.835	0.054	-2.77	0.006	-0.178
	X_{13}	Polytechnic or SS	0.994	0.046	-0.14	0.891	0.160**
University size	X_{14}	Medium	0.853	0.040	-3.39	0.001	-0.037
	X_{15}	Big	0.727	0.033	-7.02	0.000	0.037
University location	X_{16}	South	1.067	0.046	1.52	0.129	0.098*
	X_{17}	Center	1.192	0.049	4.28	0.000	0.111**
	X_{18}	North_East	1.259	0.053	5.48	0.000	0.040
	X_{19}	North_West	1.410	0.059	8.17	0.000	-0.122**
Period	X_{20}	2014-2017	1.508	0.033	19.05	0.000	
		Number of obs		63035			
		LR chi2(28)		7763.18			
		Prob > chi2		0.0000			
		Log likelihood		-28474.1			
		Pseudo R^2		0.120			

Number of obs=63035; LR chi2(28)=7763.2; Prob > chi2=0.000; Log likelihood=-28474.1; Pseudo $R^2=0.120$

*Statistical significance: *p-value <0.10, **p-value <0.05, ***p-value <0.01*

† Interaction effect between the variable and the 'Period', indicating if and to what extent the weight of the variable changes in the second period

Confirming previous literature (Weerasinghe & Dedunu, 2020; Tartari & Salter, 2015; Calvo, Fernández-López, & Rodeiro-Pazos, 2019), among the individual characteristics, gender has a non-marginal effect: women show a lower propensity to collaborate with private companies (-11.8%). The last column of Table 4 indicates that such propensity for women tends also to decrease in the second period: the coefficient of the interaction effect with the "Period" variable is negative (-0.070), although statistically not significant. Age as well shows a systematically negative impact on the response variable of the proposed model, with all odds ratios below 1, decreasing over time for older cohorts (over 45 aged), at least according to the sign of the interaction coefficient.

As for the academic rank, it shows a positive effect on the propensity for collaborating with industry, slightly decreasing in the second period: full and associate professors show a higher propensity to collaborate as compared to assistant professors, respectively by 80.9 percent and 35.4 percent.

Regarding the level of research diversification/specialization, a highly diversified scientific profile has 34.6 percent higher probability of collaborating with companies than an "intermediate" profile. Specularly, the dual covariant (highly specialized profile) has a negative impact on the propensity to collaborate (-38%).

Finally, the effect of what can be considered an exposure variable (the number of total publications), although statistically significant, is very limited in both periods (odds ratio 1.012 in 2010-2013, and 1.010 in 2014-2017).

Among all covariates under examination, “Peers collaborating” (X_{11}) presents the highest odds ratio. Confirming the indication of Tartari, Perkmann, and Salter (2014) the presence of colleagues from the same field, actively collaborating with companies, is the most important driver for an academic, even though in the second period its marginal effect substantially decreases. This is shown by the negative and significant sign of the interaction effect in the last column (-0.172). Among contextual covariates, the trait “public” of a university has a positive effect on the propensity of faculty to collaborate, probably because financial resources for research are lower than in private universities. Furthermore, also size matters: all others being equal, in medium and large-sized university the faculty propensity to collaborate is, respectively, 14.7 and 27.3 percent lower than in small-sized ones. As for geographic localization, the coefficients of the four dummies considered indicate an increase in the propensity to collaborate with latitude, i.e. towards northern Italy, where the concentration of private R&D is the highest. For academic researchers, interaction with companies depends on the level of concentration of industrial activities in general and, in particular, of knowledge-intensive industry. As for the time pattern, in the second period the propensity to engage in cross-sector collaborations seems to significantly increase in the south and the center, while it decreases in the north-west (relatively to professors on staff at universities located in the islands).

Conclusions

In the so called knowledge-based economy we live in, knowledge is increasingly becoming a distinctive competence to achieve the competitive advantage of companies, and to sustain the economic growth of nations. Universities, as the loci of new knowledge creation, have undergone in the last decades an increasing pressure to devote a special care to the technology transfer function (Todorovic, McNaughton, & Guild, 2011; Etzkowitz & Leydesdorff, 1995; Etzkowitz, 1983), which has become their third mission, alongside the traditional ones of education and research (Etzkowitz & Leydesdorff, 1995). When we refer to public-private technology transfer we generally think the exploitation by industry of results of research conducted at universities and public research organizations. A quicker and more effective way to pursue the same results is the co-creation of new knowledge through joint public-private research projects. To that aim, a number of incentive schemes have been put into place (Davenport, Davies, & Grimes, 1998; Debackere & Veugelers, 2005).

Understanding the drivers underlying the academics’ propensity to engage in collaboration with industry is critical in formulating and targeting such incentive systems to maximize their effectiveness, especially in such countries as Italy where the current levels of collaboration are suboptimal. Also important is to be aware of how the relative weight of such drivers vary over time, especially when setting performance indicators and reward systems, or when assessing the effectiveness of the policy initiatives.

Disincentives are at play too, first of all transactions costs (Belkhodja & Landry, 2005; Drejer & Jørgensen, 2005), which grow with the cultural and cognitive distance of members of the university-industry research team (Abramo, D’Angelo, Di Costa, & Solazzi, 2011). Furthermore, it has been demonstrated that on average public-private co-authored publications have lower impact than intra-university co-authored publications, and represent a lower share of highly-cited publications (Abramo, D’Angelo, & Di Costa, 2020). Because academics are increasingly subject to evaluation of their scientific activity, this awareness might constitute an additional deterrent.

In this study, we show that in the Italian academia, the share of professors collaborating with industry raised from 17.2 percent in 2010-2013 to 25.5 percent in 2014-2017. This increase is explained partly by a higher propensity to collaborate with industry by incumbents, and partly by the higher propensity of new hired professors as compared to their older colleagues who retired. This evidence can be partly due to the increasing emphasis on the importance of the university's so-called third mission. Researchers are experiencing increasing pressure to open up their research agendas to the needs of the local and national production system. There are no yet specific incentives put in place by the policy maker in this sense, but in the current national evaluation exercise, the entrepreneurial and technology transfer activities of universities are going to be assessed for the first time, in view of their possible use in the performance-based research funding scheme adopted by the Ministry of Universities and Research. At the same time, it is also possible that the increasing recourse to collaboration with private organizations has been caused by the progressive thinning of the resources made available to public research institutions, because of the budget constraints due to the tough economic situation in Italy.

The typical profile of the academic collaborating with industry is unchanged: it is a male, full professors, relatively young, with highly diversified research activity, operating within a disciplinary team made of colleagues with a high propensity themselves to collaborate with industry, and on staff in a public university located in northwestern Italy.

The relative importance of the drivers of academics' engagement in research collaboration with industry remains quite stable. The variations of the marginal effects of such personal traits as gender, age, research diversification and academic rank are quite modest and statistically not significant.

With regard to contextual drivers, the variable showing the highest weight, i.e. the presence of peers with high propensity to collaborate with industry, undergoes a significant reduction. University localization gains weight as well, while the gap between north and south shrinks.

In interpreting results, caution is recommended due to the intrinsic limits of all bibliometric approaches to the analyse of cross-sector collaborations, not all research collaborations lead to an indexed publication, and not all joint co-authored publications reflect a real public-private collaboration.

Follow on research might concern i) the measurement of publications' impact to verify whether there occurs a trade-off between the increase of the frequency of collaborations and the quality of their outputs; ii) a field level analysis to verify to what extent results vary across fields; and iii) the extension of the analysis to broader time periods with panel data.

Future research might also inquire into the geographic proximity effect on the intensity of cross-sector research collaborations, as compared with intra-sector collaborations. Finally, the methodology can be easily applied to other nations, whereby natives can easily distinguish and reconcile public and private affiliations. This would allow cross-country comparisons for a better understanding of the phenomenon.

References

Abramo, G., D'Angelo, C. A., & Murgia, G. (2013). The collaboration behaviours of scientists in Italy: A field level analysis. *Journal of Informetrics*, *7*(2), 442–454. DOI: 10.1016/j.joi.2013.01.009

Abramo, G., D'Angelo, C.A., & Di Costa, F. (2017). Specialization versus diversification in research activities: the extent, intensity and relatedness of field diversification by individual scientists. *Scientometrics*, 112(3), 1403-1418. DOI: 10.1007/s11192-017-2426-7

Abramo, G., D'Angelo, C.A., & Di Costa, F. (2019). Authorship analysis of specialized vs diversified research output. *Journal of Informetrics,* 13(2), 564-573. DOI: 10.1016/j.joi.2019.03.004

Abramo, G., D'Angelo, C.A., & Di Costa, F. (2020). The relative impact of private research on scientific advancement. Working paper, http://arxiv.org/abs/2012.04908, last accessed on 16 April

2021.

Abramo, G., D'Angelo, C.A., & Rosati, F. (2016). A methodology to measure the effectiveness of academic recruitment and turnover. *Journal of Informetrics*, 10(1), 31-42. DOI: 10.1016/j.joi.2015.10.004

Abramo, G., D'Angelo, C.A., Di Costa, F., & Solazzi, M. (2011). The role of information asymmetry in the market for university-industry research collaboration. *The Journal of Technology Transfer,* 36(1), 84-100. DOI: 10.1007/s10961-009-9131-5.

Balconi, M., & Laboranti, A. (2006). University-industry interactions in applied research: The case of microelectronics. *Research Policy*, 35(10), 1616-1630. DOI: 10.1016/j.respol.2006.09.018

Bayer, A. E., & Smart, J. C. (1991). Career publication patterns and collaborative "styles" in American academic science. *Journal of Higher Education*, *62*(6), 613–636.

Bekkers, R., & Bodas Freitas, I. M. (2008). Analysing knowledge transfer channels between universities and industry: To what degree do sectors also matter? *Research Policy, 37*(10), 1837-1853. DOI: 10.1016/j.respol.2008.07.007

Belkhodja, O., & Landry, R. (2005). The Triple Helix collaboration: Why do researchers collaborate with industry and the government? What are the factors influencing the perceived barriers? *5th Triple Helix Conference*, 1–48.

Berbegal-Mirabent, J., Sánchez García, J. L., & Ribeiro-Soriano, D. E. (2015). University-industry partnerships for the provision of R&D services. *Journal of Business Research, 68*(7), 1407-1413. DOI: 10.1016/j.jbusres.2015.01.023

Boschini, A., & Sjögren, A. (2007). Is team formation gender neutral? Evidence from coauthorship patterns. *Journal of Labor Economics*, *25*(2), 325–365. DOI: 10.1086/510764

Bowyer, K., Kranenburg, C., & Dougherty, S. (2001). Edge detector evaluation using empirical ROC curves. *Computer Vision and Image Understanding, 84*(1), 77-103. DOI: 10.1006/cviu.2001.0931

Bozeman, B., & Gaughan, M. (2011). How do men and women differ in research collaborations? An analysis of the collaborative motives and strategies of academic researchers. *Research Policy*, *40*(10), 1393–1402. DOI: 10.1016/j.respol.2011.07.002

Calvo, N., Fernández-López, S., & Rodeiro-Pazos, D. (2019). Is university-industry collaboration biased by sex criteria? *Knowledge Management Research and Practice, 17*(4), 408-420. DOI: 10.1080/14778238.2018.1557024

D'Angelo, C. A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in large-scale bibliometric databases. *Journal of the American Society for Information Science and Technology*, 62(2), 257-69. DOI: 10.1002/asi.21460.

D'Este, P., & Patel, P. (2007). University-industry linkages in the UK: What are the factors underlying the variety of interactions with industry? *Research Policy*, *36*(9), 1295–1313. DOI: 10.1016/j.respol.2007.05.002

Davenport, S., Davies, J., & Grimes, C. (1998). Collaborative research programmes: building trust from difference. *Technovation*, *19*(1), 31–40. DOI: 10.1016/S0166-4972(98)00083-2.

Debackere, K., & Veugelers, R. (2005). The role of academic technology transfer organizations in improving industry science links. *Research Policy*, *34*(3), 321–342. https://doi.org/10.1016/j.respol.2004.12.003

Di Gregorio, D., Shane, S. (2003). Why do some universities generate more start-ups than others? *Research Policy*, 32(2), 209–227. DOI: 10.1016/S0048-7333(02)00097-5

Etzkowitz, H. (1983). Entrepreneurial scientists and entrepreneurial universities in American academic science. *Minerva*, *21*(2–3), 198–233. https://doi.org/10.1007/BF01097964.

Etzkowitz, H., & Leydesdorff, L. (1995). The Triple Helix: University - Industry - Government Relations A Laboratory for Knowledge Based Economic Development. *EASST Review*, *14*, 14–19.

Fan, X., Yang, X., & Chen, L. (2015). Diversified resources and academic influence: patterns of university–industry collaboration in Chinese research-oriented universities. *Scientometrics*, *104*(2), 489–509. DOI: 10.1007/s11192-015-1618-2

Garcia, R., Araújo, V., Mascarini, S., Santos, E. G., & Costa, A. R. (2020). How long-term university-industry collaboration shapes the academic productivity of research groups. *Innovation: Organization and Management, 22*(1), 56-70. DOI: 10.1080/14479338.2019.1632711

Giuri, P., Munari, F., Scandura, A., & Toschi, L. (2019). The strategic orientation of universities in knowledge transfer activities. *Technological Forecasting and Social Change, 138*, 261-278. DOI:

10.1016/j.techfore.2018.09.030

He, Z., Geng, X., & Campbell-Hunt, C. (2009). Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a New Zealand university. *Research Policy, 38*(2), 306-317. DOI: 10.1016/j.respol.2008.11.011

Iorio, R., Labory, S., & Rentocchini, F. (2017). The importance of pro-social behaviour for the breadth and depth of knowledge transfer activities: An analysis of italian academic scientists. *Research Policy, 46*(2), 497-509. DOI:10.1016/j.respol.2016.12.003

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18. DOI: 10.1016/S0048-7333(96)00917-1

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science, 35*(5), 673-702. DOI: 10.1177/0306312705052359

Llopis, O., Sánchez-Barrioluengo, M., Olmos-Peñuela, J., & Castro-Martínez, E. (2018). Scientists' engagement in knowledge transfer and exchange: Individual factors, variety of mechanisms and users. *Science and Public Policy, 45*(6), 790-803. DOI: 10.1093/scipol/scy020

Mansfield, E. (1995). Academic research underlying industrial innovations: sources, characteristics, and financing. *Review of Economics and Statistics*, 77(1), 55-65

Mansfield, E., & Lee, J.Y. (1996). The modern university: contributor to industrial innovation and recipient of industrial R&D support. *Research Policy*, 25 (7), 1047-1058. DOI: 10.1016/S0048-7333(96)00893-1

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., . . . Sobrero, M. (2013). Academic engagement and commercialisation: A review of the literature on university-industry relations. *Research Policy, 42*(2), 423-442. DOI: 10.1016/j.respol.2012.09.007

Schartinger, D., Schibany, A., & Gassler, H. (2001). Interactive relations between university and firms: empirical evidence for Austria. *Journal of Technology Transfer*, 26, 255–268.

Shane, S. (2004). Encouraging university entrepreneurship? The effect of the Bayh-Dole Act on university patenting in the United States. *Journal of Business Venturing*, *19*(1), 127–151. DOI: 10.1016/S0883-9026(02)00114-3

Tartari, V., & Salter, A. (2015). The engagement gap: Exploring gender differences in university - industry collaboration activities. *Research Policy, 44*(6), 1176-1191. DOI: 10.1016/j.respol.2015.01.014

Tartari, V., Perkmann, M., & Salter, A. (2014). In good company: The influence of peers on industry engagement by academic scientists. *Research Policy, 43*(7), 1189-1203. DOI:10.1016/j.respol.2014.02.003

Thune, T., Reymert, I., Gulbrandsen, M., & Aamodt, P. O. (2016). Universities and external engagement activities: Particular profiles for particular universities? *Science and Public Policy, 43*(6), 774-786. DOI: 10.1093/scipol/scw01.

Todorovic, Z. W., McNaughton, R. B., & Guild, P. (2011). ENTRE-U: An entrepreneurial orientation scale for universities. *Technovation*, *31*(2–3), 128–137. https://doi.org/10.1016/j.technovation.2010.10.009

Traoré, N., & Landry, R. (1997). On the determinants of scientists' collaboration. *Science Communication*, *19*(2), 124–140. DOI: 10.1177/1075547097019002002

Ubfal, D., & Maffioli, A. (2011). The impact of funding on research collaboration: Evidence from a developing country. *Research Policy*, *40*(9), 1269–1279. DOI: 10.1016/j.respol.2011.05.023

Weerasinghe, I. M. S., & Dedunu, H. H. (2020). Contribution of academics to university–industry knowledge exchange: A study of open *innovation* in Sri Lankan universities. *Industry and Higher Education,* DOI: 10.1177/0950422220964363

Zhao, Z., Broström, A., & Cai, J. (2020). Promoting academic engagement: University context and individual characteristics. *Journal of Technology Transfer, 45*(1), 304-337. DOI: 10.1007/s10961-018-9680-6

Gendered impact of COVID-19 pandemic on research production: a cross-country analysis based on bioRxiv

Giovanni Abramo[1], Ciriaco Andrea D'Angelo[2] and Ida Mele[3]

[1] *giovanni.abramo@iasi.cnr.it*
Laboratory for Studies in Research Evaluation, Institute for System Analysis and Computer Science (IASI-CNR), National Research Council of Italy (Italy)

[2] *dangelo@dii.uniroma2.it*
Department of Engineering and Management, University of Rome "Tor Vergata" (Italy)

[3] *ida.mele@iasi.cnr.it*
Laboratory for Studies in Research Evaluation, Institute for System Analysis and Computer Science (IASI-CNR), National Research Council of Italy (Italy)

Abstract

The massive shock of the COVID-19 pandemic is already showing its negative effects on economies around the world, unprecedented in recent history. COVID-19 infections and containment measures have caused a general slowdown in research and new knowledge production. Because of the link between R&D spending and economic growth, it is to be expected then that a slowdown in research activities will slow in turn the global recovery from the pandemic. Many recent studies also claim an uneven impact on scientific production across gender. In this paper, we investigate the phenomenon across countries, analysing preprint depositions in bioRxiv. Differently from other works, that compare the number of preprint depositions before and after the pandemic outbreak, we analyse the depositions trends across geographical areas, and contrast after-pandemic depositions with expected ones. Differently from common belief and initial evidence, in few countries female scientists increased their scientific output while males plunged.

Introduction

Following the COVID-19 outbreak in China and the Far East first, Italy and Europe shortly after, and finally the Americas, governments adopted a body of emergency measures to contrast the pandemic diffusion. Among others, mobility restrictions and social distancing caused simultaneous disruptions to both supply and demand in a globalized world economy. On the supply side, reduction of labour supply because of infections, business closures and slowdown of production because of lockdowns and social distancing, caused a decrease in production. On the demand side, notwithstanding social safety nets introduced by governments, layoffs, loss of income, and worsened economic prospects caused a reduction in household consumption and private investment.

A rapidly growing number of studies investigate the macroeconomic effects of COVID-19 pandemic across countries, sectors in individual countries, as well as on a global scale (Pagano et al., 2020; Ludvigson et al., 2020; Baqaee & Farhi, 2020; McKibbin & Fernando, 2020). According to the World Bank, the massive shock of the COVID-19 pandemic and lockdown measures to contain it have plunged the global economy into the worst economic depression since World War II.[1] The negative effect on economies around the world is expected to lead to a decline in per capita income in about 90% of countries in 2020 (Djankov & Panizza, 2020), not to even mention the long-term social effects.

Due to containment measures, research activities, both public and private, have undergone a general slowdown as well, especially in those disciplines where the presence at work and close interaction with colleagues are necessary. "Levels of self-perceived productivity dropped,

[1] https://www.worldbank.org/en/news/press-release/2020/06/08/covid-19-to-plunge-global-economy-into-worst-recession-since-world-war-ii, last accessed 25 January 2021.

where dry lab scientists were much more likely to continue carrying out their work from home as expected (29% of dry lab scientists, but only 10% of wet lab scientists, reported “at least 80% productivity”)” (Korbel & Stegle, 2020). At many major research universities, non-essential research was halted, “in what amounts to an unprecedented stoppage of academic science in modern memory” (Redden, 2020).

It is widely accepted in economic theory that R&D spending can lead to rates of return well above those expected on standard capital investment (Deleidi et al., 2019). It is to be expected then that a slowdown in research output will slow in turn global recovery from COVID-19.

In this work, we undertake empirical analysis of the impact of COVID-19 pandemic on worldwide research production in life sciences, across macro-geographical areas and distinguishing by gender. We expect that the extent of slowdown in research activities varies across countries and over time, depending on the spread of infections, the extent of social restrictions, and timing of both. Furthermore, the adoption of smart working, especially at universities and public research institutions, alongside the shutdown of schools, caused a considerable increase in the scientists’ workload at home that could impact research production differently across gender. In fact, the more extensive involvement of women in family responsibilities, mainly care for children (Schiebinger & Gilmartin, 2010), might have increased or relieved because of the presence of men at home. Specularly, men might face more distractions and an intensification of domestic responsibilities when confined to the home.

The implications of findings are twofold. First, any forecasts of the impact of research on economic recovery and growth, might be misleading if based on R&D spending during the pandemic. In fact, COVID-19 pandemic did not affect so much overall R&D spending, at least by governments (Radecki & Schonfeld, 2020), while it did affect research productivity, as we will show empirically. Second, if the pandemic unevenly affects research productivity across gender, any research performance evaluation should account for that, in order not to disfavour either gender in their careers and access to resources, and institutions with uneven gender distributions of research staff.

The first empirical studies on the effects of COVID-19 pandemic on research activities focused on the response from researchers to address health issues to minimize its impact. While findings need to be verified at a later stage of the pandemic and in the years to come, from these very first investigations we learn that the volume of publications for this topic noticeably increased (Zang et al., 2020), while their quality seems below the quality average of other articles in the same journals (Zdravkovic, Berger-Estilita, Zdravkovic, & Berger, 2020). Differently from previous publications though, there seems to be a high degree of convergence between articles shared in the social web and citation counts (Kousha & Thelwall, 2020).

Soon after, the focus of scholars extended to investigate the effect of COVID-19 also on scientists’ behaviour and on research activities other than COVID-19-related research. In terms of scientists’ research behaviour, it is evident that the pandemic emergency led to substantial innovation in research collaboration and scholarly communication. The sense of urgency that has pervaded the world scientific community has generated amounts of data sharing and scientific research collaboration at levels that have never been seen before. It has been reported a speed-up of open early-stage research sharing, with a surge of depositions to such preprint archives as medRxiv and bioRxiv to foster large-scale early-stage research communication (Callaway, 2020).

The question of whether pandemic is disproportionately hurting the productivity of female scholars has been posed and empirically confirmed in different research disciplines and from different data sources: among Italian astronomy and astrophysics researchers (Deleidi et al., 2020); among neuro-immunologists (Ribarovska et al., 2021); among corresponding authors in medRxiv, but not in bioRxiv (Wehner, Li, & Nead, 2020); in the physical-sciences repository arXiv and, contrary to Wehner, Li and Nead (2020), in bioRxiv as well for the life sciences

(Frederickson, 2020); in 11 preprint repositories, expanding disciplinary coverage, and especially on COVID-19-related research (Vincent-Lamarre, Sugimoto, & Larivière, 2020). Also, early journal submission data suggest that COVID-19 is disproportionally plunging women's research production.[2]

Given that submission data are not publicly available, in the present study we recur to preprint repositories as data sources. In the past few years, there has been a significant uptake in posting preprint works in such repositories, to accelerate the diffusion of new knowledge. Considering this, and differently from other studies on the subject, we measure the variation in research production during the pandemic, by comparing research production in the pandemic period with the expected one, as extrapolated from the trends, rather than with that in previous period.

The paper unfolds as follows. In the next section we present methods and data. In the third section we present the results of the empirical analysis. The final section is devoted to the discussion and conclusions.

Data and methods

Initially, we have examined the appropriateness of a number of preprint archives to meet our objectives and methodology. We discarded all of them but arXiv or bioRxiv. The reason was that either they are too recent to be able to extrapolate the trends of posting, or the volume of preprints is too small to assure robust elaborations. arXiv is a free online archive and distribution service for unpublished preprints primarily for research in physics, astronomy and mathematics. bioRxiv is the equivalent in the life sciences. We finally chose bioRxiv as the only data source, because arXiv provides authors' affiliations of a very small share of preprints (around three percent), which makes it unsuitable for country-level analyses like the one we intend to conduct.

bioRxiv was launched in 2013 by Cold Spring Harbor Laboratory, a not-for-profit research and educational institution,[3] recently obtaining support by the Chan-Zuckerberg initiative (Callaway, 2017). First depositions occurred in November 2013.

Articles are not peer-reviewed before being posted online on this archive but undergo a basic screening process for checking non-scientific content and plagiarism. Generally, an article may be posted prior to, or concurrently with, submission to a journal so that the authors are able to make their findings immediately available to the scientific community and receive feedback on draft manuscripts.

Before the pandemic outbreak, very few bibliometric studies had used this platform as a data source. Tsunoda, Sun, Nishizawa, Liu and Amano (2019) investigated the evolution of a set of papers posted on bioRxiv and then published in academic journals. Among others, the authors found that only around 40 percent of 2013 - 2019 depositions are then published in peer-reviewed scientific journals. Fraser, Momeni, Mayr and Peters (2019) investigated the citation and altmetric advantage of depositing preprints to bioRxiv. Kenekayoro (2020) recently showed that although the platform is not yet mature enough for reliable analyses, the exponential growth in preprint depositions suggests that this data source will be soon a valuable resource for discovering interesting trends on emerging or dying research fronts.

To observe the effects of the COVID-19 pandemic and of the consequent containment measures on the production of novel scientific knowledge at a short temporal distance from its outbreak, biorXiv appears particularly appropriate. Even assuming editors' acceptance rates unchanged, the use of a traditional bibliometric platforms (such as Scopus, WoS, Google Scholar,

[2] https://www.insidehighered.com/news/2020/04/21/early-journal-submission-data-suggest-covid-19-tanking-womens-research-productivity, last accessed 25 January 2021

[3] https://www.biorxiv.org/about-biorxiv

Dimension or the like) would in fact require a longer time window considering the average publication time of an article in a journal, and its indexing in bibliographic repertories.

Moreover, bioRxiv provides free and unrestricted access to all preprints posted on the server. This applies also to machine analysis of the content. Metadata is made available via a number of dedicated RSS feeds and APIs resources. For the purpose of this research, we used a wget script for retrieving all publication metadata in XML format. Data extraction took place on 16 December 2020. After retrieving all XML files related to the original deposition (version 1)[4] of all preprints on the bioRxiv server, we implemented a parser in Python for extracting relevant information from each XML file. More in detail, we extracted: date of deposition, doi (record digital identifier in bioRxiv), author first names and last name, author position, corresponding author, institution name and country, and subject area.

Since the full set of processed XML files from bioRxiv is deposited each month with delivery completing typically in the first days of the subsequent month, we can be confident that the dataset contains all depositions up to 31 November 2020.

The full retrieved dataset is made of 106,050 preprints, showing a quadratic growth along years up to the 2020 pandemic: 77 depositions in 2013, 848 in 2014, 1704 in 2015, 4590 in 2016, 11191 in 2017, 20512 in 2018, 29018 in 2019, and 38110 in 2020. The overall distribution by bioRxiv subject area is shown in Table 1.

For the identification of the gender of each author we queried the “Gender API” platform[5] by the “first name”+”affiliation_country” pair. Since the level of standardization of bioRxiv retrieved data was not very high, some initial effort was needed for manually cleaning and reconciling fields, mainly for removing umlauts and other non-ASCII characters in first names as well as for reconciling country names.

Table 1. Share of total bioRxiv depositions by subject area

Subject area	Share	Subject area	Share
Neuroscience	17.8%	Immunology	3.2%
Bioinformatics	9.1%	Plant Biology	3.2%
Microbiology	9.0%	Developmental Biology	3.0%
Genomics	6.1%	Systems Biology	2.5%
Evolutionary Biology	5.9%	Bioengineering	2.4%
Cell Biology	5.2%	Animal Behavior and Cognition	1.5%
Genetics	4.9%	Physiology	1.4%
Biophysics	4.4%	Epidemiology	1.4%
Ecology	4.4%	Pharmacology and Toxicology	1.0%
Cancer Biology	3.6%	Synthetic Biology	0.9%
Molecular Biology	3.5%	Scientific Communication and Education	0.7%
Biochemistry	3.5%	Other	1.3%

Results

Spatiotemporal analysis of the pandemic impact

The bioRxiv monthly depositions in Figure 1 fit a quadratic trend up to the end of spring 2020,[6] with a peak in June 2020. After that, we observe an abrupt inversion of the curve, indicating a disruptive effect of the pandemic on research production.

[4] Authors can deposit a revised version of an article at any time prior to its formal acceptance by a journal.

[5] https://gender-api.com/ last accessed 25 January 2021

[6] The diagram shows the quarterly moving averages, to filter chance fluctuations.

To better appreciate the spatiotemporal dynamics of depositions before and after the pandemic outbreak, we stratify the data by geographical area. To start with, we identify the corresponding author of each preprint, the relevant affiliation and the corresponding geographical macro-area.

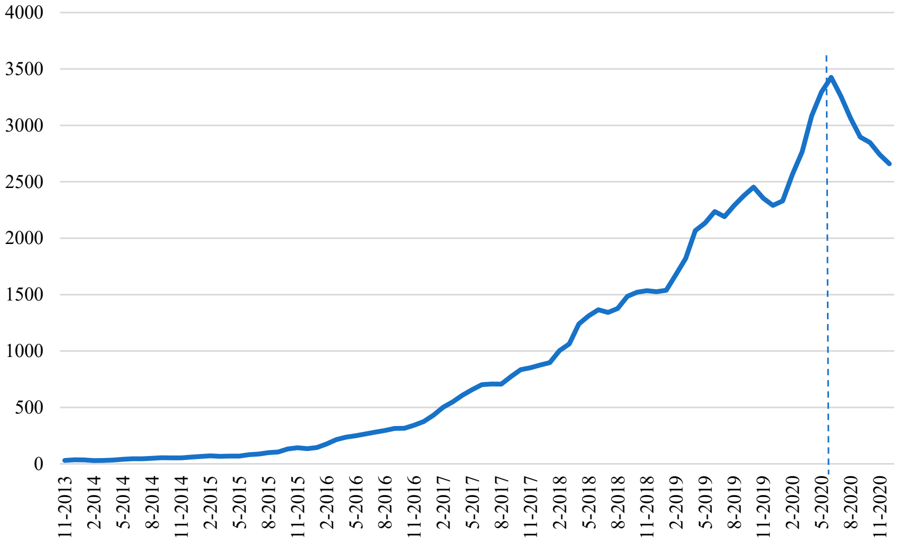


Figure 1. Time series of overall bioRxiv preprint depositions

Out of 106,050 preprints, 97,891 (92.3%) show at least one corresponding author (124,593 in total, as few publications have more than one). 101,647 corresponding authors (81.6% out of total) are provided with an affiliation[7], which can be unequivocally localized in a country and then in a geographical macro-area.

Figure 2 shows the plots of the time series depositions in three macro-areas: Europe, North America (including Canada, USA, and Greenland) and Far East (including China, Japan, Korea and Taiwan). The dot lines represent the quadratic interpolation of the yearly moving average, as measured from November 2013 to April 2020 for the Far East, and to June 2020 for Europe and North America. In order to better visualize the curve inversion in the last period, the plot starts from 2016.

The abrupt inversion of the trend is more noticeable in Europe and North America than in the Far East. The time series reflect the timing of pandemic outbreak in the different geographical areas, occurring first in the Far East but with relatively weaker effects.

It is important to notice that the average number of corresponding authors per preprint hardly changes after the pandemic outbreak (1.27 up to June 2020; 1.29 afterwards. Therefore, all trends observed with reference to the corresponding authorships remain valid for the preprints as well.

[7] Corresponding authors with multiple affiliations are counted multiple times (155,369 total affiliations).

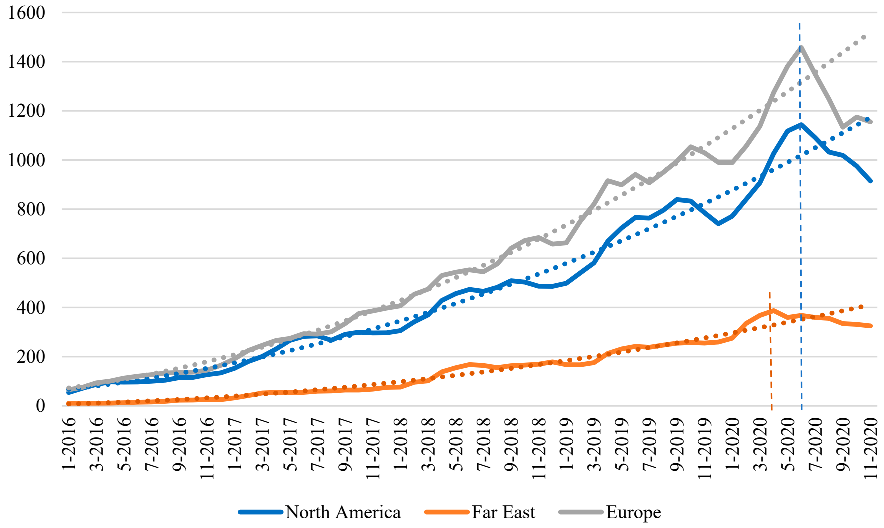


Figure 2. Time series of bioRxiv preprint depositions by macro area of the corresponding author (2016-2020 data)

Table 2 shows the percentage variations between observed and expected number of depositions in the period from September to November 2020.[8] The expected number of depositions is calculated as the product of the trend by the monthly seasonality coefficients derived from the time series.[9] We observe a drop in the number of depositions vìs-a-vìs the expected values of -17% at world level, with a maximum in Europe (-21%) and a minimum in the Far East (-8.8%).

Table 2. Observed and expected bioRxiv preprint depositions by macro-area of affiliation of the corresponding authors (September-November 2020 data)

	Observed	Expected	Variation	Confidence interval*
North America	2927	3356	-12.8%	[-9.1%;+15.5%]
Far East	994	1090	-8.8%	[-10.2%;+26.7%]
Europe	3523	4461	-21.0%	[-5.3%;+7.4%]
World	8543	10288	-17.0%	[-4.0%;+7.4%]

** Min and max monthly variations recorded for September-November 2014-2019 data*

The recorded variation is outside the confidence interval[10] at world level, for Europe and North America, but not for the Far East.

If we consider the first author in place of the corresponding author, we obtain the plot of Figure 3, showing a pattern similar to Figure 2, the only difference being the absolute values, because in a publication there is only one first author, but there could be more than one corresponding author.

[8] We chose this latest subperiod, because it is less affected by the inertia of depositions related to research projects started well in advance of the pandemic outbreak.

[9] With respect to the world level trend, every month presents average systematic variations, ranging from a minimum of -11.6% in December to a maximum of +5.5% in March in the overall period observed.

[10] Given by the min-max error of the estimate, i.e. the min-max variation between observed and expected values for September-November time series data.

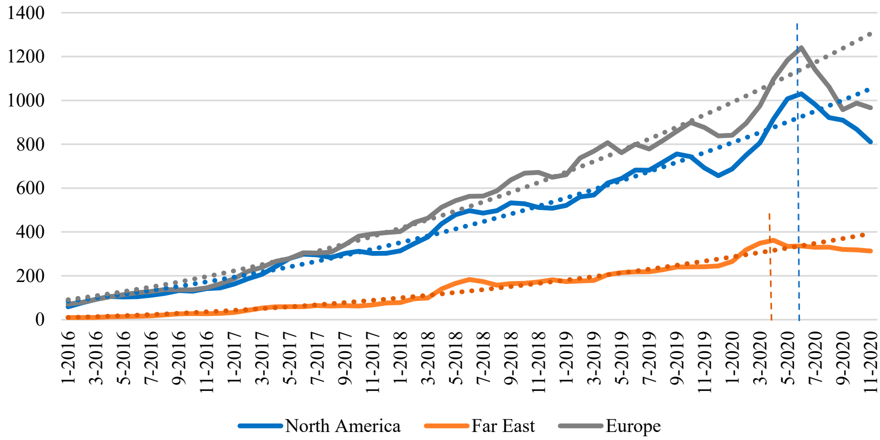


Figure 3. Time series of bioRxiv preprint depositions by macro-area of affiliation of the first author (2016-2020 data)

A gender analysis of COVID-19 impact on research production

We repeat the previous analysis further stratifying the dataset by country and gender. To discriminate gender, the first name of the author is needed alongside the country of affiliation. We discard all corresponding authors to which the Gender-API is not able to associate a gender with an accuracy equal or above 90 percent.[11] The dataset for this analysis consists of 77,156 corresponding authors, representing 61.8 percent of 24,593 total corresponding authors in bioRxiv, and 75.9 percent of those (101,647) are provided with an affiliation that can be unequivocally localized in a country.

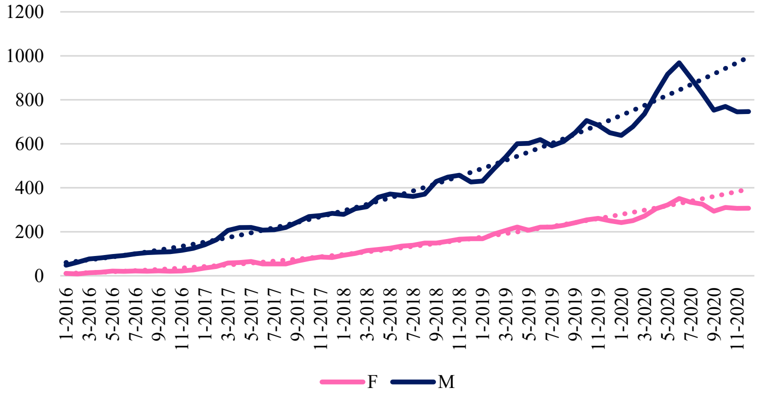


Figure 4. Time series of bioRxiv preprint depositions by gender of corresponding authors affiliated to European research organizations

[11] The authors are aware of the low accuracy of Gender_API in gender prediction of Asian and, in particular, Chinese names (Zhao & Kamareddine, 2018). However, it is the application of the threshold (90%) on the accuracy of the gender prediction to limit the extent of errors and distortion of results. In fact, 56.4% of Chinese corresponding authors have been assigned a gender and are included in the analysis as compared to 74.7% of the whole population.

Figures 4 to 6 show the time series for preprints posted by corresponding authors affiliated respectively to European, North American and Far Eastern research organizations. The reader should not be misled by the eye: when comparing the "expected" and "observed" curves around the pandemic outbreak, it seems that the plunge in production in Europe and in the Far East is much more severe for men than for women, while it is more or less the same in North America.

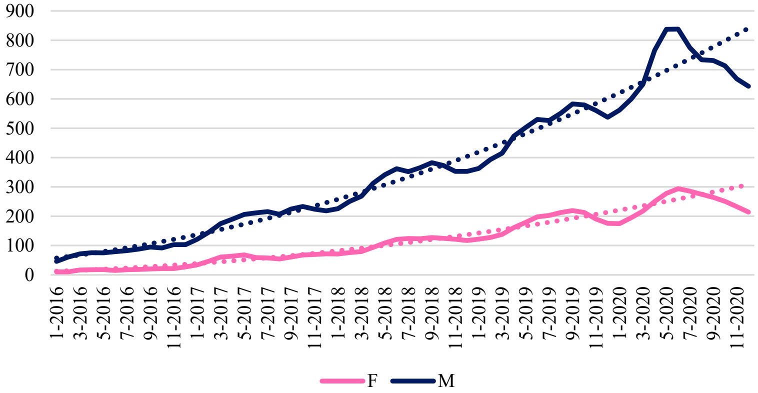


Figure 5. Time series of bioRxiv preprint depositions by gender of corresponding authors affiliated to North American research organizations

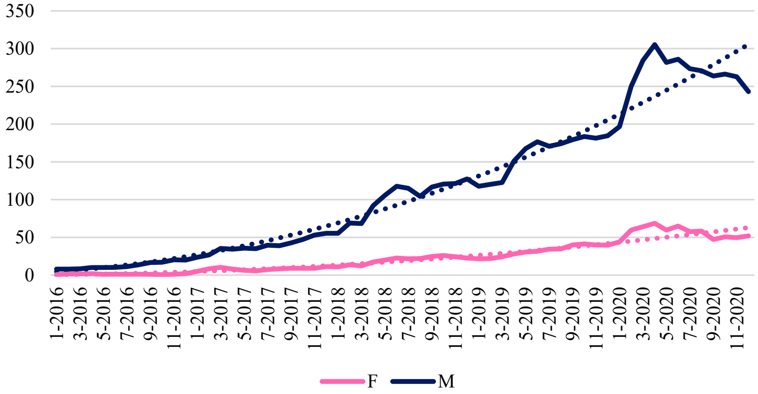


Figure 6. Time series of bioRxiv preprint depositions by gender of corresponding authors affiliated to Far East research organizations

A more in-depth analysis though, which takes into account also the seasonality of depositions, reveals that for North America (Table 3) the plunge in depositions is more conspicuous for women (-12.0%) than for men (-8.9%), similarly to the Far East (-6% vs +1.6%). In Europe instead, it is men who experienced a worse decrease in depositions (-18.8% vs -17.0%).

Table 3: Observed and expected bioRxiv preprint depositions by macro-area and gender of corresponding authors (September-November 2020 data)

	Gender	Observed	Expected	Variation
North America	F	752	855	-12.0%
	M	2139	2347	-8.9%
Far East	F	152	162	-6.0%
	M	799	786	+1.6%
Europe	F	931	1121	-17.0%
	M	2310	2845	-18.8%

We delve the analysis at country level considering the U.S., China, and the top eight European countries per number of depositions. Findings in Table 4 confirm the unbalance showed in Table 3 in the U.S., where female scientists reduced the output of their research activities at a higher rate than their male colleagues, -13.6% vs -9.3%. The same occurs in China, but to a less extent: -5.5% for females vs -3.0% for males. In Europe, contrasting evidences emerge. In France, Italy, Netherlands, and Switzerland it is women who are hurt more, while in Germany and Spain the opposite holds true. Quite surprisingly, in both countries female scientists raised their depositions with respect to the expected ones, respectively by 10 percent and 45 percent, while males decreased theirs by 15 percent and 9 percent. In Sweden and U.K., gender differences are hardly noticeable. It must be said that the more fine-grained the analysis the less robust the results, because of the lower number of observations. Nevertheless, what emerges at continental level is often untrue at country level, where the interplay of different containment measures, women's role in society, and family-related infrastructure unveil quite different realities.

Table 4: Observed and expected bioRxiv preprint depositions by gender of the corresponding authors in the U.S., China, and the top eight European countries per number of depositions (September-November 2020 data)

Country	Gender	Observed	Expected	Variation
United States	F	665	769	-13.6%
	M	1919	2117	-9.3%
China	F	111	117	-5.5%
	M	402	414	-3.0%
France	F	123	168	-26.6%
	M	276	344	-19.8%
Germany	F	199	181	+9.9%
	M	509	596	-14.6%
Italy	F	34	47	-28.4%
	M	98	110	-10.8%
Netherlands	F	39	60	-35.5%
	M	136	147	-7.3%
Spain	F	73	50	+45.4%
	M	127	140	-9.1%
Sweden	F	32	45	-28.5%
	M	64	92	-30.6%
Switzerland	F	47	65	-27.7%
	M	184	188	-2.1%
United Kingdom	F	214	309	-30.7%
	M	507	734	-31.0%

Conclusions

The human impacts of COVID-19 infections, and pandemic-related limitations and impediments are vast. These include disruptions to researchers that, as we showed, differ by nation and gender. In this work, we have investigated the impact of pandemic on research output. Results confirm what was expected that is a general plunge in preprint depositions, following the pandemic outbreak, consistent in timing across geographical areas: in China and the Far East first, in Europe and North America then. Probably, we will never see the consequences of the slowdown in scientific production caused by COVID-19 pandemic, in terms of published articles, as journal editors can simply raise their acceptance rates to keep the volumes unchanged. Most likely, we will witness lower average quality of publications. Evidence of that can be assessed in the years to come. Editors though might give now a precious contribution to scholars in the field by providing them with data on submission variations after the pandemic outbreak.

Contrary to what most people might expect, and early studies have announced, that female scientists are hurt more by pandemic due to the increase in family care workload, we could observe that this holds true at world level while significant exceptions occur at country level. The important lesson to be learnt is that world level analyses often hide significant differences across countries, especially when country-specific variables play a significant role in determining the outcomes.

Why gender differences occur across countries would require further investigation by scholars knowledgeable about the single country under observation. As for Italy, results are not surprising to us. In Italy women's relative share of involvement in family responsibilities, mainly care for children but also for parents and parents-in-law, is more extensive than in average EU countries. In Italy, for example, only 24 percent of the children go to kindergarten,[12] not allowing parents to be full-time occupied (Istat, 2019). Percent of population ages 65 and older (often associated with high levels of morbidity) is 24 in Italy, among the highest in the world (PRB, 2020). According to a recent survey, the majority of Italians totally agreed on the statement: "The most important role of a woman is to take care of her home and family" (EU, 2017).

Findings might be of interest to scholars in scientometrics, in the economics of innovation, and in sociology. The pandemic effects on research can inform policy makers when dealing with economic forecasts, gender equality issues, research evaluation exercises, and the assessment of the effectiveness of relevant policies and initiatives.

We appreciate several limitations embedded in our study. The field of analysis is limited to the life sciences, therefore findings and conclusions cannot be generalized to other disciplines. Data extraction was conducted during the pandemic, whose expiration is hopefully expected to occur in the next few months. Therefore, the extent of the effects that we tried to grasp is to be confirmed by future updates. It is the intention of the authors to extend the period of investigation to June 2021, in order to provide the ISSI conference participants with up-to-date and more robust findings. We also intend to extend the analysis to all authors of publications.

Finally, future research might entail investigation on the pandemic impact on research collaboration behaviour. It is to be expected in fact that intra-muros collaborations must have lost their advantage over extra-muros, as physical presence and personal contacts were inhibited by containment measures.

References

Baqaee, D.R., & Farhi, E. (2020). Nonlinear production networks with an application to the COVID-19 crisis. *CEPR,* discussion paper 14742.

Callaway, E. (2017). bioRxiv preprint server gets cash boost from Chan Zuckerberg initiative. *Nature*, 545(7652), 18.

Callaway, E. (2020). Will the pandemic permanently alter scientific publishing? *Nature*, *Science After the Pandemic*, 582, 167–68.

[12] Pandemic containment measures spared kindergarten, which remained open most of the pandemic period.

Capelle-Blancard, G., & Desroziers A. (2020). The stock market and the economy: Insights from the COVID-19 crisis, *VoxEU.org*, 19 June.

Deleidi, M., De Lipsis, V., Mazzucato, M., Ryan-Collins, J., & Agnolucci, P. (2019). *The macroeconomic impact of government innovation policies: A quantitative assessment*. UCL Institute for Innovation and Public Purpose, Policy Report working paper series (IIPP WP 2019-06). https://www.ucl.ac.uk/bartlett/public-purpose/wp2019-06, last accessed 16 April 2021.

Djankov, S., & Panizza, U. (2020). COVID-19 in Developing Economies, a VoxEU.org eBook, CEPR Press.

EU (2017). Special Eurobarometer 465: Gender Equality 2017. *EU Open Data Portal*, https://data.europa.eu/euodp/en/data/dataset/S2154_87_4_465_ENG, last accessed 16 April 2021.

Fraser, N., Momeni, F., Mayr, P., & Peters, I. (2019). Examining the citation and altmetric advantage of bioRxiv preprints. Paper presented at the *17th International Conference on Scientometrics and Informetrics, ISSI 2019 - Proceedings, 1,* 667-672

Frederickson, M. (2020). COVID-19's gendered impact on academic productivity (GitHub, 2020) https://github.com/drfreder/ pandemic-pub-bias, last accessed 16 April 2021.

Inno, L., Rotundi, A., & Piccialli, A. (2020). COVID-19 lockdown effects on gender inequality. *Nature Astronomy,* 4(12), 1114.

Istat (2019). *Asili nido e altri servizi socio-educativi per l'infanzia. Report*, available at: https://www.istat.it/it/files/2019/03/asili-nido.pdf, last accessed 16 April 2021.

Kenekayoro, P. (2020). Author and keyword bursts as indicators for the identification of emerging or dying research trends. *Journal of Scientometric Research, 9*(2), 120-126.

Korbel, J.O., & Stegle, O. (2020). Effects of the COVID-19 pandemic on life scientists. *Genome Biology,* 21(113), 1-5, https://doi.org/10.1186/s13059-020-02031-1.

Kousha, K., & Thelwall, M. (2020). COVID-19 publications: Database coverage, citations, readers, tweets, news, Facebook walls, Reddit posts. Quantitative Science Studies, 1(3), 1068–1091.

Lee, J.J., & Haupt, J.P. (2020). Scientific globalism during a global crisis: Research collaboration and open access publications on COVID-19. *Higher Education,* doi:10.1007/s10734-020-00589-0

McKibbin, W.J. & Fernando, R. (2020). The global macroeconomic impacts of COVID-19: seven scenarios, *CAMA,* working paper 19/2020.

Pagano, M., Wagner, C., & Zechner, J., (2020). COVID-19, asset prices, and the great reallocation, *VoxEU.org*, 11 June.

PRB-Population Reference Bureau (2020). *Countries with the oldest population in the world. Report-* Available at https://www.prb.org/countries-with-the-oldest-populations/, last accessed 16 April 2021.

Radecki, J., & Schonfeld, R. C. (2020). *The Impacts of COVID-19 on the Research Enterprise: A Landscape Review*. https://doi.org/10.18665/sr.314247, last accessed 16 April 2021.

Redden, E. (2020). Empty benches at empty lab tables. *Inside Higher Ed*, https://www.insidehighered.com/news/2020/03/30/nonessential-research-has-halted-many-campuses, last accessed 16 April 2021.

Ribarovska, A. K., Hutchinson, M. R., Pittman, Q. J., Pariante, C., & Spencer, S. J. (2021). Gender inequality in publishing during the COVID-19 pandemic. *Brain, behavior, and immunity*, *91*, 1–3.

Schiebinger, L. & Gilmartin, S.K. (2020). Housework is an academic issue. *Academe* 96, 39–44.

Tsunoda, H., Sun, Y., Nishizawa, M., Liu, X., & Amano, K. (2019). An analysis of published journals for papers posted on bioRXiv. *Proceedings of the Association for Information Science and Technology, 56*(1), 783-784.

Vincent-Lamarre, P., Sugimoto, C.R., & Larivière V. (2020). The decline of women's research production during the coronavirus pandemic. *Nature Index*. https://www.natureindex.com/news-blog/decline-women-scientist-research-publishing-production-coronavirus-pandemic, last accessed 16 April 2021

Wehner, M. R., Li, Y., & Nead, K. T. (2020). Comparison of the proportions of female and male corresponding authors in preprint research repositories before and during the COVID-19 pandemic. *JAMA network open*, *3*(9), e2020335.

Woolston, C., (2020). Pandemic darkens postdocs' work and career hopes. *Nature*, Work, 585, 309-12.

Zdravkovic, M., Berger-Estilita, J., Zdravkovic, B., & Berger D. (2020). Scientific quality of COVID-19 and SARS CoV-2 publications in the highest impact medical journals during the early phase of the pandemic: A case control study. *PLoS ONE*, 15(11): e0241826.

Zhang, L., Zhao, W., Sun, B., Huang, Y., & Glänzel, W. (2020). How scientific research reacts to international public health emergencies: A global analysis of response patterns. *Scientometrics,* 124(1), 747-773.

Zhao, H., & Kamareddine, F. (2018). Advance gender prediction tool of first names and its use in analysing gender disparity in computer science in the UK, Malaysia and China. *Proceedings of the 2017 International Conference on Computational Science and Computational Intelligence, CSCI* 2017, 222-227. DOI:10.1109/CSCI.2017.35

Research Unit Size and Internal Co-publications

Hind Achachi[1] and Hamid Bouabid[2]

[1] *hindachachi@gmail.com*
Ibn Tofail university, B.P 242, Kénitra (Maroc)

[2] *h.bouabid@fsr.ac.ma*; *h.bouabid@hotmail.com*
Mohammed V University of Rabat, Avenue Ibn Batouta
Agdal, BP1014 RP (Maroc)

Abstract

Scientific collaboration within a research unit (team, laboratory, group, etc.) has caught scientometricians' attention. Science of team science recently emerged to use science to understand the best ways researchers are jointly organized to perform science in order to increase their performance as well as to suggest relevant strategies to support effective science teams. This paper contributes to this field of science by analyzing to which extent internal collaboration outputs of a research unit is influenced by its size, i.e. number of academics. Therefore, 120 accredited units composed of about 1,500 researchers located within four universities in Morocco are analyzed in this paper. A cross-matrix of members' co-publications is built for each unit. An index of intra-collaboration and an internal co-publication indicator are calculated to assess both size and internal research output at the unit's level. Furthermore, the Lorenz curve is being used to draw equality/inequality of the internal co-publication with respect to unit's size. The results indicate low internal co-publications for larger research units with an inequality in the distribution of these publications in favor of small size units.

Introduction

The analysis of scientific collaboration allows putting forward patterns, organizing modes, obstacles and opportunities for effective and impactful research at all levels of collaboration. Scientific collaboration represents a social component of modern science (Glanzel and Schubert, 2005; Wuchty et al., 2007; Milojevic, 2010). It is observed on several levels: micro (individual), meso (unit, institution, etc), and macro (national, international, field, etc). Since the publication of the first works by Hooke *et al.* in 1665 on the topic (see Beaver & Rosen, 1978), it is recently a matter of science of team science (Hall et al., 2018; Liu et al., 2020).

The complexity of team collaboration requires adopting appropriate tools to measure and evaluate targeted specific collaboration' facets. As pointed out by Glanzel and Schubert (2005) and Melin and Persson (1996), co-authorship is a particular outcome of collaboration. It can be used as a proxy to study collaboration. Persson et al. (2004), Wuchty et al. (2007) and Larivière et al. (2014), all found that co-authored publications were more frequently cited than sole publications, which results in increasing outputs' impact and quality. Furthermore, these co-publications have a positive impact on scientific productivity at the individual, institutional, and national levels, as well as on socio-economic partnerhips (Helga et al., 2009, Lebeau et al., 2008). This strong correlation is independent of the context (Lee and Bozeman, 2005, Haslam et al, 2008, Defazio et al, 2009, Abramo et al., 2011, Bouabid, 2014).

In this rich literature on scientific collaboration, the majority of studies is particularly oriented towards inter-institutional, international, national collaborations or in specific scientific disciplines (Kumar, 2015; Savic et al., 2015; Wuchty et al. 2007; Jones et al., 2008). Intra-institutional collaboration has also caught academic attention. In a comprehensive literature review by Kumar (2015), two such studies were conducted by Newman (2004) and by Pepe and Rodriguez (2010). Bellanca (2009), De Stefano et al. (2011) and Birnholtz et al. (2013) analyzed intra-institutional collaboration of the University of York (UK), University of Salerno (Italy), and two campuses of Cornell University (USA) respectively.

Besides addressing research performance at a micro-level as well as at a macro one, a number of researchers studied it at the meso-level: research unit or laboratory, as the first 'cell' for scientific collaboration. Indeed, the research unit is put forward as the basic unit for its role as a site of idea emergence, knowledge creation, diffusion and discovery (Bonaccorsi and Daraio, 2005; Carayol and Matt, 2004; Von Tunzelmann et al., 2003; Horta and Lacy, 2011).

All these works at research unit level as the one by Sandstrom and Van den Besselaar (2019) examined the effect of the unit size on the unit's whole output (papers), productivity and impact, without distinguishing internal co-publications. Partitioning the outputs to bring internal co-publications out is crucial because being formally organized together in a unit, members are supposed to team up for collaborative outputs. Indeed, as advocated by Sandstrom and Van den Besselaar (2019), the performance - productivity and citation impact - of a research group relies on its capacity to combine and use different competencies in a creative way.

Bringing out internal co-publications at the research unit is the first contribution of this paper, introducing in this regard a new quantitative method based on a purposely cross-matrix shaping internal co-authorship. Its second contribution is that it addresses intra-collaboration at a unit's level within the context of a developing country, which very few researches have dealt with. Indeed, except the research by Aparecido et Kannebley (2019) in Brazil - to the best of our knowledge - all the others were conducted in developed countries (Bonaccorsi et al. (2006) in an Italian context, Carayol and Matt (2006), in a French context, Cook et al. (2015), in a UK context, Sandstrom and Van den Besselaar (2019), in a European context, Brandt and Schubert (2013), in a German context, Verbree et al. (2015), in a Dutch context, De Saá-Pérez et al. (2017) and in a Spanish context. Shin et al. (2013) found that collaboration patterns differ across systems: researchers in developed countries are more collaborative than their peers in developing ones.

For greater performance and visibility of university research, the Ministry of Scientific Research in Morocco launched an initiative in 2006 for the structuration of the scientific landscape. This structuration constitutes the first milestone towards reaching critical unit size and then research performance. Thus, universities have adopted a common national platform that sets criteria for organizational research units: (i) *Team* of at least 3 professors, (ii) *Laboratory*, of at least 3 research teams or 9 professors, and (iii) a *Center* as a group of teams and laboratories. The process of structuration has allowed the transition to a more formal system where research units are necessarily accredited by the University Council based on prior-defined eligibility criteria. After at least two successive accreditations of four years each, does this initiative have a positive impact on intra-collaboration output?

Data and methods

This research analyzes the unit's intra-collaboration of academics being part of accredited research units after a four-year period following their accreditation cycle of the period 2008-2011. It covers the Faculties of Sciences (*FS*) performing research in the fields of science, technology, engineering and mathematics, within four Moroccan universities out of twelve[1]:

- Mohammed V University in Rabat (*UMV*) ;
- Ibn Tofail University in Kenitra (*UIT*) ;
- Mohammed First University in Oujda (*UMP*) ;
- Moulay Ismail University in Meknès (*UMI*).

The present research focuses only on *Laboratories* as research units, to which we refer later as unit, since the *Team* unit (an average of 3 academics in all institutions under study) is not large enough to allow an objective and rational analysis of intra-collaboration.

Our sample included these four universities composed of 120 research laboratories out of a total of 175, and 1,500 academics out of a total of 1,701, which accounted respectively for 36% and

[1] The three main criteria for this choice among other universities were age, size and geographical location.

33% of all universities in the field of science, technology, engineering and mathematics (see table 1). Moreover, these four universities count for 44% of the total PhD fellows at universities. Descriptive information is given in table 1 regarding the weights of the 4 universities in terms of the numbers of research units, academics (full professors), and PhD fellows.

Table 1: Number of research units, academics and PhD fellows in the 4 universities under this study, and in all universities with their evolution over a decade from 2009 to 2018. Values in parentheses indicate the share of the 4 universities out of all universities

		2009	2018	Change (18/09)
Research units : all universities	Team	445	326	-27 (%)
	Laboratory	488	690	41 (%)
	Center	20	51	155 (%)
	Other	29	0	
	Total	982	1067	9 (%)
Research units : sample (4 universities)	Team	116 (26%)	120 (37%)	3 (%)
	Laboratory	175 (36%)	156 (23%)	-11(%)
	Center	0 (0%)	25 (49%)	
	Other	8 (28%)	0	
	Total	299 (30%)	301 (28%)	1 (%)
Academics: all universities	STEM	5037	6628	32 (%)
	MED	1216	1532	26 (%)
	SSH	3850	5794	50 (%)
	Total	10,103	13,954	38 (%)
Academics: Sample (4 universities)	STEM	1,680 (33%)	2,479 (37%)	48 (%)
	MED	21 (2%)	714 (47%)	3,300 (%)
	SSH	1,020 (26%)	1647 (28%)	61 (%)
	Total	2,721 (27%)	4,840 (35%)	78 (%)
PhDs fellows: all universities	STEM	6,970	14,352	106 (%)
	MED	1,599	3,084	93 (%)
	SSH	10,279	16,877	64 (%)
	Total	18,848	34,313	82 (%)
PhDs fellows: Sample (4 universities)	STEM	3,084 (44%)	5,195 (36%)	68 (%)
	MED	0 (0%)	1,206 (39%)	
	SSH	4,159 (40%)	6,199 (37%)	49 (%)
	Total	7,243 (38%)	12,600 (37%)	74 (%)

STEM: Science, Technology, Engineering and Mathematics, MED: Medical sciences, SSH: Social science and Humanities

In terms of disciplines, the fields of physics and environment were the most represented compared to other scientific fields (Figure 1). It is worth informing that all research units are mainly disciplinary-focused, with an organization more linked to departmental structures (which are teaching units) at the university, except for the environment' field.

After identifying the accredited units in the different disciplines, we proceeded to retrieving the name and surname of each academic in these laboratories. Then, we listed the academics' names in a particular manner in a matrix, which we call "intra-collaboration matrix", as shown in Figure 2. For example, for a unit with *n* academics, the intra-collaboration matrix is built by placing the academics' names in rows, from bottom to top, numbered from *1* to *n-1,* then crossing them by placing the academics' names this time in columns from left to right, academics' names starting by the n^{th} remaining academic and counting down to academic number *2*. The matrix allows crossing all unit's academics to plot their respective internal co-authored and sole papers. The intra-collaboration matrix is symmetrical.

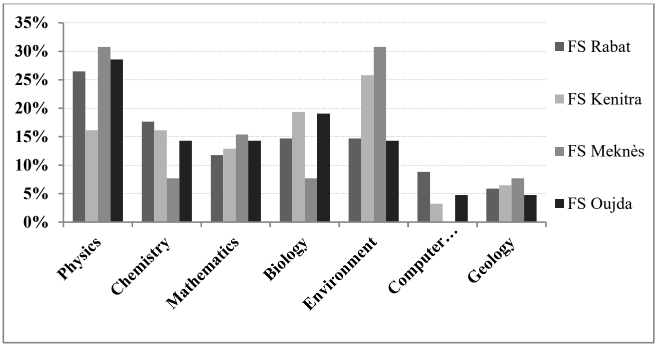
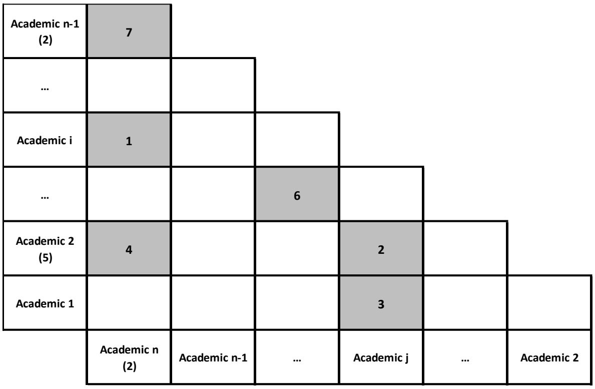


Figure 1: Distribution of accredited research units according to scientific disciplines

To read the matrix in figure 2, as an example, the academic *2* has 4 co-authored publications with academic *n* and 2 with academic *j*. The numbers between parentheses in the cells, in the left column and the lower row, refer to sole publications of the academics in the unit. A sole publication refers here to a publication made by an academic without any co-authoring with their unit members, which may be a single authored or a co-authored publication with a third party outside the unit. As an example, academic *2* has 5 publications alone without any co-authorship within her/his research unit, besides her/his 6 internal co-publications.



Academic n-1 (2)	7					
…						
Academic i	1					
…			6			
Academic 2 (5)	4			2		
Academic 1				3		
	Academic n (2)	Academic n-1	…	Academic j	…	Academic 2

Figure 2: Example of intra-collaboration matrix for a unit with *n* academics

In this matrix, gray cells represent the number of co-publications of the same unit and number of co-authors. As an example, the academic *2* has 2 co-authors (academic n and academic j), 6 co-authored publications and 5 soles papers.

This analysis uses all types of publication in the Web of Science database (Clarivate Analytics) over a four-year period after the accreditation cycle period of 2008-2011. The four-year timespan ensures the elimination of the yearly fluctuations in the publishing process. Each publication of the corpus had the affiliation of at least one of the studied institutions. The data was retrieved in a format with all the fields of the database records. After refining the raw data, the corpus is about 1,500 publications, containing the exact affiliations, names and surnames of the academics belonging to these research units. Out of which, 49.4% are internal co-publications and 50.6% are sole publications.

Results and discussion

Research unit size measure

To quantify the research unit size, we suggest the grouping index that refers to the average number of academics in a unit within a scientific field or an institution. This index is calculated for each institution (faculty) and for each field. It is defined as the sum of the number of academics per unit for all units, divided by the number of units in a given institution or a field:

$$I_g = \frac{\sum_{i=1}^{m} n_i}{m}$$

where m: number of research units in an institution or a scientific field,
$n_{i:}$ number of academics in unit i.

The grouping index values for the four institutions are reported in Table 2 with respect to each scientific field. It shows that the largest grouping occurred at the Faculty of Meknes with more than 14 academics per unit, reflecting a strong research collectivity. Small unit size (less than 10 academics per unit) is found at the faculty of Rabat (an average of 9.9), knowing that the threshold required for constituting an accredited unit is nine academics. Moreover, at the Faculty of Rabat, there is a significant gap between the minimum and the maximum values of the grouping index *Ig*.

Table 2: Average values of the grouping index *Ig* of academics by university

	FS Oujda			FS Rabat			FS Kenitra			FS Meknes			Mean
	Min	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Field**
Physics	8	11.3	14	4	7.7	12	6	7.2	9	12	11.5	18	**9.1**
Mathematics & Computer Science	10	11.3	13	9	11.0	14	7	9.4	13	16	20.5	25	**11.7**
Geology	-	15	-	4	12.0	20	6	7.5	9	-	6.0	-	**10**
Chemistry	12	13.7	16	5	11.3	16	6	5.6	12	-	17.0	-	**10.3**
Biology & Environment	10	13.0	17	4	10.0	21	6	9.9	15	9	14.6	20	**11.2**
Mean		**12.4**			**9.9**			**8.5**			**14.1**		

Regarding the scientific fields, it can be noted that in the field of physics, academics are the least grouped ones, with an Index of about 9. On the contrary, mathematics and computer science is the field where academics tend to gather more in a research unit, irrespectively of the institution (almost 12 academics per unit).

Intra-collaboration measure

To explore the intra-collaboration outputs, we consider three indicators: (i) the intra-collaboration index, (ii) the number of cells with co-publications (cells in gray color in Figure 3) and finally, and (iii) the sum of the number of co-publications within these cells. Then, the Lorenz curve is used to depict both the last two indicators.

The index of intra-collaboration for each matrix is defined as the number of cells where there are co-publications (cells in gray color in Figure 4) divided by the number of cells in the upper (or lower) part of the matrix (without the diagonal). We only consider the upper or the lower part of the matrix because the co-publications matrix is always symmetrical.

In a mathematical form, we write this index as follows for the unit *i*:

$$I_i = \frac{c_i}{(n_i(n_i - 1))/2}$$

c_i: number of cells with co-publications,

n_i: number of academics in unit i.

This size-normalized index ranges from a value of *1* when all cells are in gray color, meaning that all academics have co-publications among them, i.e. full collaboration, and a value of *0* when there is no co-publication between the unit's academics, i.e. *"loneliness"*, as expressed by von Tunzelmann et al. (2003).

Figure 3 shows the intra-collaboration index according to the size of the research unit for all the studied units. We note that the intra-collaboration index is lower and decreases when the size increases. Some years later, the research structuration process of grouping more academics is not yet reaching collaborative publishing outputs. The reason is that achieving this goal requires that the unit goes through several stages in its life cycle, during which compatibility and interconnection at work, besides incentive and infrastructure, need to take place as key factors of collaboration (Hara et al., 2003). Yet, these stages require much time while the process of structuration is relatively recent. Enough time is also required to support strong individual relationships, which are found to allow the establishment of trust and contribute to a better transfer of information and ideas between researchers (Jha and Welch, 2010). Horta and Lacy (2011) have supported the evidence that greater intensity of communication between individuals and groups is a key factor for fostering creativeness and research results. Moreover, according to the survey conducted by Achachi et al (2016), the majority of researchers in a developing countries context stated that affinity is a fundamental vector to build and improve collaboration between them. This demonstrates that intra-collaboration does not immediately increase with unit size, but rather increases gradually with other factors such as relationship (affinity and compatibility), work methodology, writing style and interconnection at work, shared goals, incentives, etc.

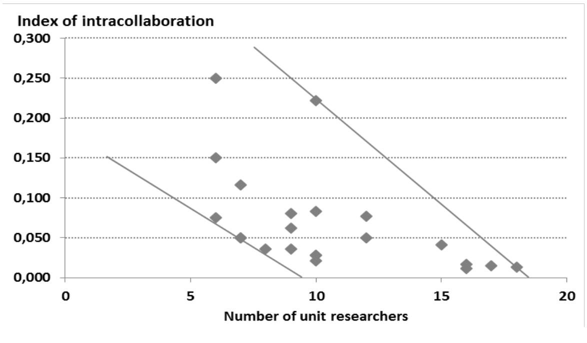


Figure 3: The intra-collaboration index with respect to the number of academics in the research units

Internal co-publications

To understand better this reverse evolution of the intra-collaboration according to the research unit size, the Lorenz curve is used to depict equality or inequality of the internal co-publications according to unit size. Figure 4 shows this relationship for all units by setting a threshold of three gray cells in the collaboration matrix to be represented. Both curves, of the number of gray cells and the total number of internal co-publications in these cells, illustrate clearly the inequality in the distribution of internal co-publications according to the size of the research unit. For total internal co-publications, 40% of the population in large-size units on average holds only 10 % of the internal co-publications (figure on the right). Nevertheless, 20% of the academics - all in small-size units - holds 50% of these papers (figure on the right). The inequality is higher when counting total internal co-publications than the number of gray cells. This finding corroborates previous empirical evidences in developed country contexts, that is size does not matter much for a unit performance (Von Tunzelmann, 2003). It may even have a negative effect on productivity regardless of the context (Qurashi, 1991, in an Indian context; Seglen and Aksnes, 2000, in a Norwegian context; Bonaccorsi et al., 2006, in an Italian context; Cook et al., 2015, in a UK context; Sandstrom and Van den Besselaar, 2019, in a European context). At best, productivity evolves in an inverted-U curve with size, up to an optimal size - between 5 and 9 members - then decreases again (Qurashi, 1993; Von Tunzelman et al., 2003; Verbree et al. 2015; De Saá-Pérez et al. 2017).

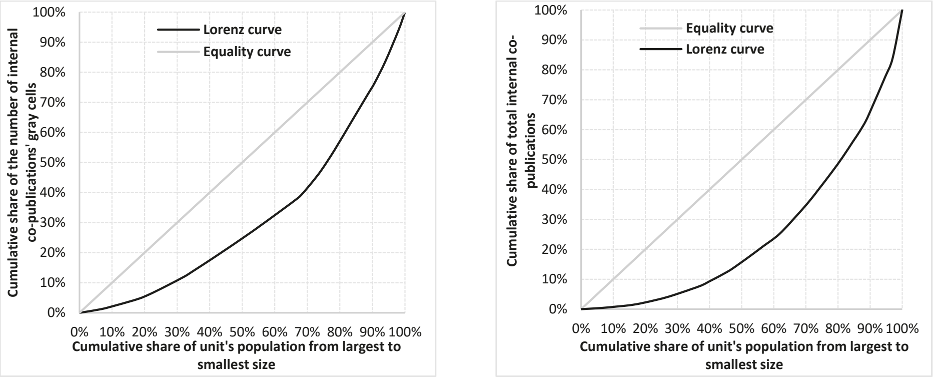


Figure 4: Distribution of Cumulative internal co-publications as a function of the cumulative unit's size (number of academics)

The observed low internal co-publication may also be because research units are built on a field specialization basis with mainly disciplinary focus. Indeed, all units under study were devoted to a specific scientific field, where unit members had similar specialization, called vertical specialization. While this scheme of unit organization seems to favor collaborative work among its members (Laudel, 2002), in the Moroccan context, it is likely supporting competitive attitude among them as they are working on the same - or almost the same - research topics. This competitive attitude is enhanced by the career promotion requirement. Perhaps, horizontal specialization, i.e. multi- or inter-disciplinary research activities, may be suitable for these research units to perform better. Indeed, Schubert (2014) put forward that "even if a certain specialization strategy is globally optimal, it does not need to be optimal for the individual research group".

Despite the academics grouping initiative implemented in universities since 2006, internal co-publications aren't found to increase with research unit's size. The main obstacles challenging research units are the lack of funding, common logistic support, 'teamwork' culture, visibility and impact of the academics' work. As a result, academics are likely to prefer other forms of collaboration, particularly international collaboration, that bring substantial funding and more recognition and visibility. The international collaboration is often developed through individual initiatives rather than formal and institutional policy or agreements (Bouabid 2017, in developing country; Bordons et al. 2013, in developed country). Likewise, Waast (2010) and Achachi et al. (2016) stated that international collaboration is preferred because it is seen to help in keeping the scientific level of researchers up-to-date, but also providing important financial support and scientific recognition.

Two other major factors appear as shortcomings to the internal unit's publishing are: (i) the lack of postdoctoral position in the university landscape and (ii) the teaching overload. With regard to the first shortcoming, academics at university in the Moroccan context are all full Professors, under three categories, from the highest: Professor of the Higher Education, Professor Habilitated for research (Associate Professor) and Assistant Professor. Postdoctoral position is not permitted by law at Moroccan universities, while this category of research personnel plays a key role in the research unit outputs (Carayol and Matt, 2006) besides both academics and PhD fellows. Furthermore, Horta and Lacy (2011) concluded that units' output is positively impacted by the type of its organization when including both academics, postdoctoral and PhD fellows, while it is negatively impacted when this organization is being limited to academics and doctoral fellows only. Cook et al. (2015) provided additional evidence that post-doctorates are more productive than PhD fellows and achieve more impact.

Regarding the second shortcoming, teaching duty, particularly in an undergraduate level, is found to reduce scientific outputs of academics. Units are typically multi-functional: mainly engaged in teaching and research activities. While Lepori et al. (2016) noticed a sort of balance of these two activities in most university units they studied, Horta et Lacy (2011) advocated that academics should be relieved from exclusively teaching undergraduates, since it negatively impacts their productivity. Expectedly, internal co-publications in Moroccan units, are low because the teaching duty is very high. Indeed, with a ratio of students to teaching staff in the faculties of science of 31.3 (and 54.5 at the whole university) in 2017, it is too far behind those of developed countries such as (in the same year): France (16.1), Germany (11.7), Italy (20.9), Netherlands (14.6), Portugal (14.7), Spain (11.7), USA (15.3).

Conclusion

This paper brings some evidence on the extent to which the research unit size at universities affects its scientific internal co-publications. These empirical findings support that in the Moroccan context - a developing country context, the policy initiative, which started recently, of grouping more academics into research units, doesn't yet result in higher collaborative outputs: internal co-publications. This finding is in line with several other researches in developed countries' contexts. Using Lorenz curve on internal co-publications, this paper demonstrates that internal co-publications are not in favor of large units, suggesting that this initiative does not seem yet to stimulate collective work resulting in co-publications. The grouping initiative is likely in its early stages to build real units, which are marked by favoring information exchange and communication between its members, but not reaching yet the upward stage of improving internal collaboration outputs. Since the formal institutional organization or research is so recent, the 'culture' of affinity, interconnection and individual relationship, are not well spread yet, and requires more time to become effective.

Since internal co-publications are found to be unequal with respect to the unit size, the science policy is asked to put forward other prerequisites to reach unit performance, such as invigorating the culture of 'teamwork' in the whole research landscape and to explore optimal unit size rather than simply extending it. Science policy towards university should be clear when it comes to research vs. educational orientation, which fundamentally governs the formal institutional organization and its performance in both research and teaching.

References

Abramo, G., & D'Angelo, C. A. (2011) 'Evaluating research: From informed peer review to bibliometrics', *Scientometrics*, 87: 499-514.

Achachi, H. Amor, Z. Dahel. Mehkhancha, C. C., Cherraj, M., Bouabid, H. Selmanovic, S., Larivière, V. (2016) 'Factoring affecting researcher's collaborative patterns: A case study from Maghreb universities', *The Canadian Journal of Information and Library Science*, 40: 234-253.

Adams J., King C., Hook D., (2010) 'Global research report: Africa', *Thomson Reuters*.

Aparecido Dias A. and Kannebley Junior, S. (2019) 'Scientific productivity and patenting at the laboratory level: an analysis of Brazilian public research laboratories', *Economics of Innovation and New Technology*, 29 (2):1-21.

Beaver, D. B., & Rosen, R. (1978) 'Studies in scientific collaboration, I. The professional origins of scientific co-authorship', *Scientometrics*, 1(1): 65-84.

Bellanca, L. (2009) 'Measuring interdisciplinary research: Analysis of co-authorship for research staff at the University of York', *Bioscience Horizons*, 2(2): 99-112.

Birnholtz, J., Guha, S., Yuan, Y. C., Gay, G., & Heller, C. (2013) 'Cross-campus collaboration: A scientometricand network case study of publication activity across two campuses of a single institution', *Journal of the American Society for Information Science and Technology*, 64(1): 162–172.

Bonaccorsi A., Daraio C., Simar L. (2006) 'Advanced indicators of productivity of universities: an application of robust nonparametric methods to Italian data', *Scientometrics*, 66(2): 389-410.

Bordons, M., Aparicio, J., and Costas,R. (2013) 'Heterogeneity of Collaboration And Its Relationship with Research Impact in a Biomedical Field', *Scientometrics*, 96 (2): 443–66.

Bouabid, H. (2014) 'Science and technology metrics for research policy evaluation: some insights from a Moroccan experience', *Scientometrics*, 101: 899-915.

Bouabid, H. (2017) 'A scientometric method for assessing an institution scientific collaboration policy', *16th International Society of Scientometrics and Informetrics (ISSI) Conference*, 856-868.

Brandt T. and Schubert T. (2013) 'Is the university model an organizational necessity? Scale and agglomeration effects in science', *Scientometrics,* 94:541–565.

Carayol N. and Matt M. (2004) 'Does research organization influence academic production? Laboratory level evidence from a large European university', *Research Policy*, 33: 1081-1102.

Carayol N. and Matt M. (2006) 'Individual and collective determinants of academic scientists' productivity', *Information Economics and Policy*, 18: 55-72.

Cook I., Grange S., Eyre-Walker A. (2015) 'Research groups: How big should they be?' PeerJ, 2015(6) doi:10.7717/peerj.989.

Defazio, D., Lockett, A., & Wright, M. (2009) 'Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program', *Research Policy*, 38: 293-305.

De Saá-Pérez P, Díaz-Díaz NL, Aguiar-Díaz I, Ballestreros-Rodríquez JL (2015) 'How diversity contributes to academic research teams performance', *R&D Management*, 47(2): 165-179.

De Stefano, D., Giordano, G., Vitale, M. P. (2011) Issues in the analysis of co-authorship networks, Quality & Quantity, 45(5), 1091-1107.

Gaillard A. M., Canesse A. A., Gaillard J., Arvanitis R. (2013) 'Euro-Mediterranean science and technology collaborations: a questionnaire survey', *Options Méditerranèennes*, B n° 71: 2013.

Glänzel, W., & Schubert, A. (2005) 'Analysing scientific networks through co-authorship'. In: *Handbook of quantitative science and technology research*. pp. 257–276. Dordrecht (NL): KluwerAcademic Publishers.

Hall, K. L., Vogel, A. L., Huang, G. C., Serrano, K. J., Rice, E. L., Tsakraklides, S. P., & Fiore, S. M. (2018). The science of team science: A review of the empirical evidence and research gaps on collaboration in science. American Psychologist, 73(4), 532-548.

Hara, N., Solomon, P., Kim, L., and Sonnenwald, D. H. (2003) 'An Emerging View of Scientific Collaboration: Scientists Perspectives on Collaboration and Factors that Impact Collaboration', *Journal of the American Society for Information Science and Technology*, 54(10):952-965.

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., et al. (2008) 'What makes an article influential? Predicting impact in social and personality psychology', *Scientometrics*, 76(1): 169-185.

Helga, B. A., Ernesto, L. R. L., & Tomas, B. M. (2009) 'Dimensions of scientific collaboration and its contribution to the academic research groups scientific quality', *Research Evaluation*, 18(4): 301-311.

Horta, H., and T. A. Lacy. (2011) 'How Does Size Matter for Science? Exploring the Effects of Research Unit Size on Academics Scientific Productivity and Information Exchange Behaviors', *Science and Public Policy*. 38(6): 449-462.

Horta, H, V Dautel and F Veloso (2012). 'An output perspective on the teaching–research nexus: an analysis focusing on the US higher education system', *Studies in Higher Education*, 37(2): 171-187.

Jha, Y., and Welch, E. W., (2010) 'Relational mechanisms governing multifaceted collaborative behavior of academic scientists in six fields of science and engineering', *Research Policy* , 39: 1174-1184.

Jones B. F., Wuchty S., Uzzi B. (2008) 'Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science', *Science*, 322(5905): 1259-1262.

Landini, F., Malerba, F., Mavilia, R. (2015) 'The structure and dynamics of networks of scientific collaborations in Northern Africa', *Scientometrics*, 105 (3): 1787-1807.

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2014) 'Team size matters: Collaboration and scientific impact since 1900', *Journal of the Association for Information Science and Technology*, 66(7): 1323-1332.

Laudel, G. (2002) 'What do we measure by co-authorships?', *Research Evaluation*, 11(1): 3-15.

Lebeau, L., Laframboise, M. C., Larivière, V., & Gingras, Y. (2008) 'The effect of university–industry collaboration on the scientific impact of publications: The Canadian case, 1980-2005', *Research Evaluation*, 17(3): 227-232.

Lee, S. & Bozeman, B. (2005) 'The Impact of Research Collaboration on Scientific Productivity'. *Social Studies of Science*, 35(1): 673-702.

Lepori, B., Wise, M., Ingenhoff, D., Buhmann, A., (2016) 'The dynamics of university units as a multi-level process. Credibility cycles and resource dependencies', Scientometrics, 109 (3): 2279-2301.

Liu, Y., Wu, Y., Rousseau, S., Rousseau, D. (2020), Reflections on and a short review of the science of team science. *Scientometrics* 125, 937-950. https://doi.org/10.1007/s11192-020-03513-6

Melin, G., & Persson, O. (1996) 'Studying research collaboration using co-authorships', *Scientometrics*, 36(3): 363-377.

Milojevic´, S. (2010) 'Modes of collaboration in modern science: Beyond power laws and preferential attachment', *Journal of the Association for Information Science and Technology*, 61(7): 1410-1423.

Newman, M. E. J. (2004) 'Coauthorship networks and patterns of scientific collaboration', *Proceedings of the National Academy of Sciences*, 101(1): 5200-5205.

Pepe, A., & Rodriguez, M. A. (2010) 'Collaboration in sensor network research: An in-depth longitudinal analysis of assortative mixing patterns', *Scientometrics*, 84(3): 687-701.

Persson, O., Glänzel, W., &Danell, R. (2004) 'Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies' *Scientometrics*, 60: 421-432.

Qurashi M. M. (1991) 'Publication-rate and size of two prolific research groups in departments of inorganic-chemistry at Dacca University (1944-65) and zoology at Karachi University (1966-84)', *Scientometrics*, 20(1): 79-92.

Sandtrom U. and Van den Besselaar P. (2019) 'Performance of Research Teams: results from 107 European groups', *17th International Conference of the International-Society-for-Scientometrics-and-Informetrics (ISSI)*, Vol II, 2240-2251.

Savic´, M., Ivanovic´, M., Radovanovic´, M., Ognjanovic´, Z., Pejovic´, A., & Jaksˇic´ Krger, T. (2015) 'Exploratory analysis of communities in co-authorship networks: A case study'. In: A. M. Bogdanova & D. Gjorgjevikj (Eds.), ICT innovations 2014, *advances in intelligent systems and computing*, (Vol. 311, pp. 55–64). Berlin: Springer International Publishing.

Shin, J. C., Lee, S. J., & Kim, Y. (2013) 'Research collaboration across higher education systems: Maturity, language use, and regional differences', *Studies in Higher Education*, 38(3): 425-440.

Schubert, T. (2014). Are there scale economies in scientific production? on the topic of locally increasing returns to scale. Scientometrics, 99(2), 393-408. doi:10.1007/s11192-013-1207-1

Seglen, P.O. & Aksnes, D.W. (2000). Scientific productivity and group size: A bibliometric analysis of Norwegian microbiological research. Scientometrics, 49 (1). 125-143.

Verbree M., Horlings E., Groenewegen P., Van der Weijden I., Van den Besselaar P. (2015) 'Organizational factors influencing scholarly performance: a multivariate study of biomedical research groups', *Scientometrics*, 102: 25-49.

Von Tunzelmann N Ranga M, Martin B, Geuna A (2003) 'The Effects of Size on Research Performance: A SPRU Review', *SPRU Report*, Sussex, U.K.

Waast, R. (2010) 'Research in Arab Countries (North Africa and West Asia)', *Science, Technology & Society*, 15:2 (2010): 187-231.

Wuchty S., Jones B. F., Uzzi B. (2007), The increasing dominance of teams in production of knowledge, *Science*, 316(5827): 1036-1039.

Scholarly publishing at US federally funded research & development laboratories: influences on public-private science

Jeffrey M. Alexander[1], Cassidy R. Sugimoto[2] and Vincent Larivière[3]

[1] *jmalexander@rti.org*
RTI International, 3040 East Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC, USA

[2] *sugimoto@indiana.edu*
School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA

[3] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, 3150 rue Jean-Brillant, Montréal, QC, Canada

Abstract

Government investment in basic research is generally divided into two categories: funding to external organisations—such as universities—and funding to internal organizations. While an extensive body of literature has been devoted to universities' contribution of basic science, much less is known about that of governmental research centers. This paper analyzes the contribution of federally funded research and development centers (FFRDCs) to US science, with a focus on the contribution of the research & development laboratories (RDLs) to scholarly papers, scientific impact, disciplinary specialization and collaboration patters. Results show that these RDLs account for an increasing part of US papers—representing more than 25% of US papers in some specialties of physics—and that their relative impact is higher than that of average US papers. They are also found to vary substantially in their collaboration patterns, with percentages of international collaboration on publications from these organizations oscillating between 18% and 85%.

Introduction

Government investment in basic research produces benefits that counteract the forces that lead the private sector to underinvest in such research (Rosenberg, 2009). Governments differ in how they organize publicly funded scientific research. In general, this spending is divided between funding distributed to performers outside of the government, such as universities and private firms (i.e., extramural R&D), and funding to government research facilities (intramural R&D). National governments establish intramural research facilities, often termed "public research institutes" or PRIs, for a variety of reasons. Intarakumnerd and Goto (2018) assert that newly industrialized nations tend to establish PRIs to strengthen their national innovation systems and accelerate their pursuit of technological competitiveness, while developed nations use PRIs for a wider range of purposes. A significant motivation for creating and maintaining PRIs is to conduct research that would not normally be undertaken in industry or academia. Military research is a typical activity for many PRIs. National defense is provided almost uniquely by national governments. While some military research may be outsourced to industry and even to academia, extramural mechanisms may compromise secrecy and government influence over the direction and nature of the research performed. Another common activity of PRIs is fundamental, curiosity-driven research, especially when that research requires a significant investment in instrumentation and infrastructure. Universities and firms seldom have the incentive, resources, or expertise to maintain very large and specialized scientific instruments, such as radio telescopes and particle accelerators. Those facilities are operated commonly either by public agencies, or by extramural performers with significant funding and oversight from government.

Few studies of PRIs exist in the published literature. Recognizing this, the OECD launched a study in 1989 on government research laboratories, with a follow-up study conducted in 2011

(Sanz-Menéndez et al., 2011). In the United States, Bozeman and Crow (1998) published a series of examinations of federal government research laboratories. Cruz-Castro et al. (2020) and Zacharewicz et al. (2017) produced a series of scholarly works looking at organizational and managerial issues in public research organizations, and Hallonsten (2017) recently conducted a review of what he terms the "third sector of R&D" (encompassing government research laboratories but including non-governmental research institutes). This study focuses on a specific form of PRI in the United States: nationally owned research laboratories operated as federally funded research and development centers (FFRDCs). The FFRDC is an organizational form different from most PRIs in other nations. These institutions are typically formed and owned by the U.S. federal government, but operated by academic, industrial, and special-purpose entities under bespoke agreements with specific agencies (known as the FFRDC's "sponsor"). As a result, these PRIs straddle the boundary between intramural and extramural government laboratories. They retain the status of a governmental entity, but with many of the attributes of a non-governmental organization.

The United States spends more on research and development as a nation than any other country—estimated at $580 billion in 2018 (National Center for Science and Engineering Statistics, 2020). Of that amount, slightly below 10 percent is performed at federal government research organizations. The scale of the U.S. R&D enterprise accommodates a wide range of organizational types—especially forms of public-private partnerships. Since the 1980s, such partnerships grew more common as a mechanism for sharing and creating knowledge by leveraging the respective capabilities and perspectives of government agencies, universities, and private firms (Carayannis & Alexander, 1999; Carayannis et al., 2000). The FFRDC is an early example of this type of government-university-industry partnership. FFRDCs evolved from their origin in World War II to a significant part of the U.S. public R&D enterprise, accounting for approximately 36 percent of the research performed at federal facilities in 2018.There are three categories of FFRDCs: R&D Laboratories (RDLs), Study and Analysis Centers (SACs), and Systems Engineering and Integration Centers (SICs). Of those, (RDLs) are the major producers of multi-disciplinary scientific knowledge. Study and Analysis centers tend to produce technical reports that are highly tailored to the interests of their sponsoring government agencies, and less frequently produce more generalizable knowledge appropriate for publication in the scholarly literature. Systems Engineering and Integration centers conduct substantial applied research and develop new methods for system development and validation, but these projects are also designed to meet specific government needs. For this study, we focus on FFRDCs that are considered as R&D laboratories, as they have the workforce, facilities, and latitude to conduct quasi-academic research across the physical, social, and life sciences. This research in progress analyzes the place of FFRDCs in the US basic research landscape. We first assess the relative importance of FFRDCs in US science, from both the output and impact points of view. We then analyze their disciplinary specialization, as well as their collaboration patterns at the international level.

Methods

Data for this paper was obtained from the Web of Science (WOS) database, covering the Science Citation Index Expanded, the Social Sciences Citation Index, and the Arts and Humanities Citation Index, over the 1973-2018 period. The name variations of each 24 FFRDC analyzed were retrieved from the institution and department fields of the WOS, leading to a conversion table of 195 name variations at the level of institutions and 7,919 at the level of departments.[1] Using this conversion table, we retrieved 441,636 distinct papers (articles, notes and reviews), and 484,447 FFRDC-paper combinations (including multiple counts of papers in case of collaboration between FFRDCs). Each paper was assigned the subfield of the journal in which it is published according to the NSF field and subfield classification of journals (Hamilton, 2003), and this classification scheme was used for the field-normalization of their number of citations (Waltman et al., 2011). Full counting of papers was used and networks were created with UCINET and Netdraw.

Results

Over the 1980-2018 period, the RDLs contributed to 410,699 papers, representing 3.7 percent of all papers with a US address. This percentage has been rising thoroughly over the period, from 3.4 percent in the 1980s to 3.9% in the 2010s, even as the share of FFRDC performance of US basic research (measured by funding amount) declined from 13% in the 1980s to 5.7% in the 2010s (NCSES, 2021). RDLs vary in output, however. Four laboratories have contributed to more than 1000 papers per year on average over the last four decades: Lawrence Berkeley National Laboratory (58,872 papers), Los Alamos National Laboratory (54,209 papers), Oak Ridge National Laboratory (48,812 papers) and Argonne National Laboratory (43,911 papers). At the other end of the spectrum, others have not been as active in basic research, such as the Software Engineering Institute (33 papers retrieved) and the Center for Advanced Aviation System Development (3 papers retrieved). On average, RDLs have contributed to more than 18,000 papers throughout the period. The scientific impact (average of relative citations) of RDLs is also above the world average as well as the US average. In the 2010 decade, RDL papers have, on average, been cited twice as much as the average papers of the disciplines in which they published. The lead RDL in this regard is Sandia National Laboratories, a FFRDC sponsored by the Department of Energy’s National Nuclear Security Administration. Despite its status as a RDL focused on weapons research, Sandia illustrates the significance of basic research in supporting its sponsor’s mission (National Research Council, 2013). We also observe that the average impact of RDLs and of US papers are going in opposite directions: while that of RDLs is increasing slightly, US papers are, on average, less and less cited, from 52% above world average in the 1980s to 34% in the 2010s. The contribution of RDLs to basic research varies by domain. There are 17 specialties, however, in which they globally account for more than 10 percent of all US papers over the 2010-2018 period (Figure 1). Their contribution is remarkable in Nuclear Technology, Nuclear and Particle Physics and Solid State Physics, where they participated to 42, 31 and 25 percent, respectively, of all US papers published during the period. Other areas with an important relative contribution to US science are related to earth and space science, as well as other areas of applied physics, chemistry and engineering. The disciplinary specialization varies by RDLs, however (Figure 2). While most RDLs are active in specialities of Physics, Chemistry and Engineering and drive the aggregated trends observed in Figure 1, we observe that the Frederick National Lab for Cancer Research is, unsurprisingly, active in various fields of health sciences, and that the Software Engineering Institute is focusing on Computers. Similarly, the National Center for Atmospheric Research and the National Solar Observatory are focusing on Meteorology and Atmospheric Sciences.

[1] The cleaning of the Software Engineering Institute (SEI), however, proved to be more difficult. This FFRDC is based at Carnegie Mellon University, and researchers from the Institute often sign the name of the University without indicating their affiliation to SEI. Therefore, our results from SEI underestimate its contribution to basic research.

Table 1. Number of papers and average of relative citations of RDL FFRDCs, by decade, 1980-2018.

FFRDC	Number of papers				Average of relative citations			
	1980s	1990s	2000s	2010s	1980s	1990s	2000s	2010s
Lawrence Berkeley National Laboratory	8,081	11,034	16,192	23,565	2.13	1.95	2.24	2.45
Oak Ridge National Laboratory	8,902	10,718	12,090	17,102	1.45	1.60	1.72	1.81
Argonne National Laboratory	7,166	8,692	11,340	16,713	1.88	1.75	1.78	2.02
Los Alamos National Laboratory	9,998	12,735	16,234	15,242	1.51	1.64	1.67	1.74
Pacific Northwest National Laboratory	1,810	3,242	5,960	10,072	1.51	2.01	1.66	1.92
Jet Propulsion Laboratory	3,979	5,743	7,534	10,035	1.67	1.64	1.61	1.93
Brookhaven National Laboratory	6,430	6,552	8,104	9,876	1.80	1.85	1.89	2.01
Lawrence Livermore National Laboratory	5,435	8,402	10,336	9,467	1.88	1.80	1.71	1.93
Sandia National Laboratories	4,845	6,325	7,140	7,754	2.24	2.05	1.76	1.53
SLAC National Accelerator Laboratory	1,535	1,461	2,768	7,044	1.97	2.43	2.19	2.93
Frederick National Laboratory	4,013	6,471	6,557	6,891	2.08	1.68	1.52	1.68
National Center for Atmospheric Research	2,068	2,857	4,387	5,991	1.90	1.91	1.95	2.12
National Renewable Energy Laboratory	969	1,404	2,196	3,965	2.34	2.17	2.64	2.27
Fermi National Accelerator Laboratory	1,610	2,160	2,966	3,711	1.50	2.07	2.34	2.64
Ames Laboratory	1,541	2,368	3,235	2,901	1.65	1.62	1.66	1.41
National Radio Astronomy Observatory	902	1,245	1,635	2,114	1.74	1.44	1.34	1.79
Princeton Plasma Physics Laboratory	200	863	1,338	1,975	1.74	1.17	1.43	1.47
Idaho National Laboratory	319	987	1,260	1,862	0.85	1.00	1.18	1.28
National Optical Astronomy Observatory	426	991	1,046	1,221	1.48	1.54	2.14	2.52
Lincoln Laboratory	1,121	1,146	764	762	2.25	2.27	1.61	1.29
Thomas Jefferson National Accelerator Facility	57	422	659	756	1.08	1.78	1.49	1.30
Savannah River National Laboratory	219	209	371	619	1.03	0.67	0.69	0.77
National Solar Observatory	152	308	357	419	0.94	1.06	0.78	0.84
Software Engineering Institute	2	12	9	10	1.92	0.21	0.44	0.34
Center for Advanced Aviation System Development			2	1			0.62	0.30
All FFRDCs	69,054	90,526	113,019	138,100	1.80	1.78	1.81	2.00
% of US papers	3.4%	3.6%	3.8%	3.9%	-	-	-	-
United States	2,050,175	2,522,706	2,981,575	3,559,831	1.52	1.44	1.38	1.34

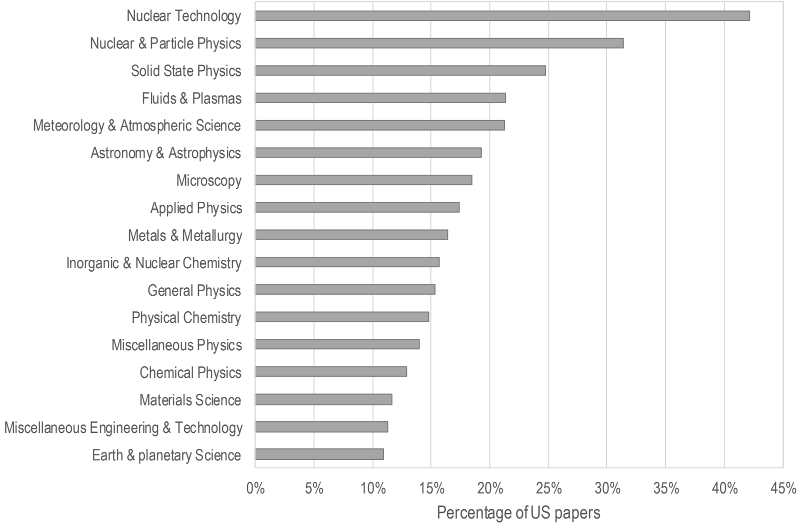


Figure 1. Specialties in which RDL FFRDCs (taken altogether) account for more than 10% of US papers, 2010-2018.

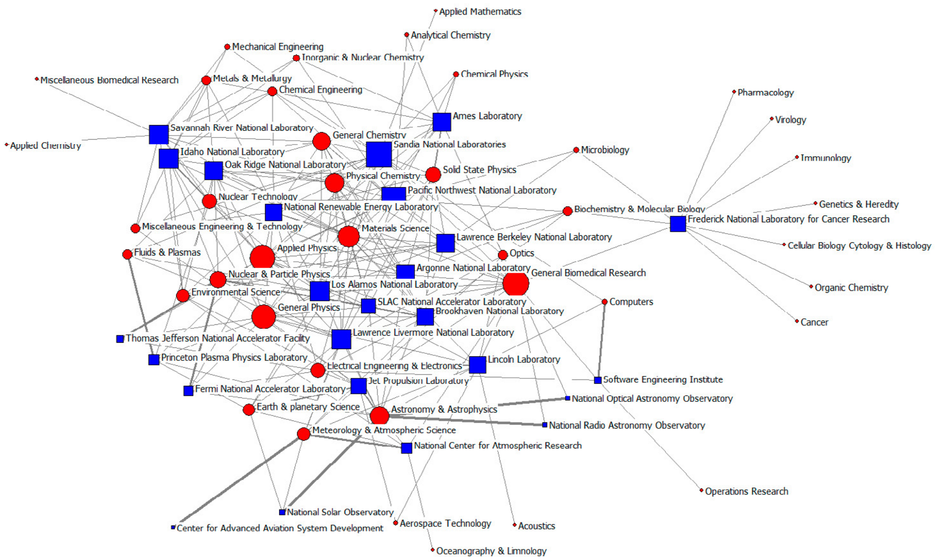

Applied Mathematics
Analytical Chemistry
Mechanical Engineering
Inorganic & Nuclear Chemistry
Miscellaneous Biomedical Research
Metals & Metallurgy
Chemical Physics
Chemical Engineering
Pharmacology
Ames Laboratory
Virology
Savannah River National Laboratory
General Chemistry
Applied Chemistry
Sandia National Laboratories
Idaho National Laboratory
Microbiology
Immunology
Oak Ridge National Laboratory
Solid State Physics
Physical Chemistry
Pacific Northwest National Laboratory
Nuclear Technology
Genetics & Heredity
National Renewable Energy Laboratory
Biochemistry & Molecular Biology
Frederick National Laboratory for Cancer Research
Miscellaneous Engineering & Technology
Materials Science
Lawrence Berkeley National Laboratory
Cellular Biology Cytology & Histology
Fluids & Plasmas
Applied Physics
Optics
Argonne National Laboratory
General Biomedical Research
Nuclear & Particle Physics
Los Alamos National Laboratory
Organic Chemistry
Environmental Science
SLAC National Accelerator Laboratory
Computers
General Physics
Brookhaven National Laboratory
Cancer
Thomas Jefferson National Accelerator Facility
Lawrence Livermore National Laboratory
Princeton Plasma Physics Laboratory
Lincoln Laboratory
Electrical Engineering & Electronics
Software Engineering Institute
Fermi National Accelerator Laboratory
Jet Propulsion Laboratory
National Optical Astronomy Observatory
Earth & planetary Science
Astronomy & Astrophysics
National Radio Astronomy Observatory
Meteorology & Atmospheric Science
National Center for Atmospheric Research
Operations Research
National Solar Observatory
Center for Advanced Aviation System Development
Aerospace Technology
Acoustics
Oceanography & Limnology

Figure 2. Network of specialization between RDLs and specialties, for specialties that represent at least 2% of a FFRDC's total papers, 2010-2018.

RDLs exhibit strong differences in their international collaboration patterns (Table 2). While the vast majority of papers from astronomical observatories are in international collaboration, this percentage is much lower for large facilities such as Lawrence Berkeley National Laboratory, Oak Ridge National Laboratory, and Argonne National Laboratory, and even lower in the cases of Lincoln Laboratory and Savanah River National Laboratory. The main international collaborator is also varying by laboratory, with Germany and China being the main collaborators of 9 and 8 laboratories, respectively.

Table 2. Percentage of papers in international collaboration and main international collaborator, by RDL, 2010-2018.

FFRDC	Top internat. collaborator	% of papers in internat. collaboration
National Radio Astronomy Observatory	Germany	84.6%
National Optical Astronomy Observatory	Germany	78.7%
Fermi National Accelerator Laboratory	UK	75.6%
SLAC National Accelerator Laboratory	Germany	70.6%
Thomas Jefferson National Accelerator Facility	Italy	64.6%
National Solar Observatory	Germany	63.0%
Jet Propulsion Laboratory	France	60.7%
Brookhaven National Laboratory	China	59.1%
Lawrence Berkeley National Laboratory	Germany	57.1%
National Center for Atmospheric Research	UK	56.5%
Argonne National Laboratory	China	54.3%
Lawrence Livermore National Laboratory	Germany	47.5%
Princeton Plasma Physics Laboratory	China	47.3%
Los Alamos National Laboratory	Germany	45.9%
Oak Ridge National Laboratory	China	45.5%
Ames Laboratory	China	45.0%
Frederick National Laboratory for Cancer Res.	China	43.1%
Pacific Northwest National Laboratory	China	37.7%
National Renewable Energy Laboratory	China	31.6%
Idaho National Laboratory	France	25.7%
Sandia National Laboratories	Germany	24.4%
Lincoln Laboratory	Germany	18.8%
Savannah River National Laboratory	Japan	17.8%

Conclusion

The contribution of RDLs to basic research in the US is sizeable and increasing. Results show that RDLs account for an increasing part of US papers—representing more than 25% of US papers in some specialties of physics—and that their relative impact is higher than that of average US papers. They are at the heart of research in nuclear physics, and contributing in an important fashion to several disciplines of earth and space sciences. Despite their strong US national component—their research focus can be considered to be of strategic importance for the country—an increasingly large proportion of their papers is co-authored with international colleagues, and Germany and China represent the largest collaborators. RDLs vary dramatically in their collaboration patterns, with percentages of international collaboration oscillating between 18% and 85%. On the whole, this research in progress paper shows that the RDL FFRDCs, which often host unique scientific instruments, have an important contribution to US science, and complement research performed in universities and industrial research centers. Further research will investigate the national and international collaboration networks of RDLs, their citation linkages—the organizations cited in RDL publications, and the organizations that cite RDL work—and the relationship between their scientific activities and the funding and priorities of their sponsoring agencies.

References

Bozeman, B. & Crow, M. (1998). *Limited by Design: R&D Laboratories in the U.S. National Innovation System*. New York, NY: Columbia University Press.

Carayannis, E. & Alexander, J. (1999). Winning by co-opeting in strategic government-university-industry R&D partnerships: the power of complex, dynamic knowledge networks. *The Journal of Technology Transfer*, 24(2–3), 197–210.

Carayannis, E. G., Alexander, J. & Ioannidis, A. (2000). Leveraging knowledge, learning, and innovation in forming strategic government–university–industry (GUI) R&D partnerships in the US, Germany, and France. *Technovation*, 20(9), 477–488.

Cruz-Castro, L., Martínez, C., Peñasco, C. & Sanz-Menéndez, L. (2020). The classification of public research organizations: Taxonomical explorations. *Research Evaluation*, 1–15. doi: 10.1093/reseval/rvaa013

Hallonsten, O. (2017). The third sector of R&D: Literature review, basic analysis, and research agenda. *Prometheus*, 35(1), 21–35.

Hamilton, K. (2003). *Subfield and level classification of journals* (CHI Report No. 2012-R). Cherry Hill, NJ: CHI Research

Intarakumnerd, P. & Goto, A. (2018). Role of public research institutes in national innovation systems in industrialized countries: The cases of Fraunhofer, NIST, CSIRO, AIST, and ITRI. *Research Policy*, 47(7), 1309–1320.

National Center for Science and Engineering Statistics (NCSES) (2021). *National Patterns of R&D Resources: 2018–19 Data Update*. NSF 21-325. Alexandria, VA: National Science Foundation. Available at https://ncses.nsf.gov/pubs/nsf21325.

National Research Council 2013. *The Quality of Science and Engineering at the NNSA National Security Laboratories.* Washington, DC: The National Academies Press.

Rosenberg, N. (1990). Why do firms do basic research (with their own money)? *Research Policy*, 19(2), 165-174.

Sanz-Menéndez, L., Cruz-Castro, L., Jonkers, K., Derrick, G. E., Bleda, M. & Martinez, C. (2011). *Policy Brief: Public Research Organisations*. Paris: OECD.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S. & van Raan, A. F. (2011). Towards a new crown indicator: An empirical analysis. *Scientometrics*, 87(3), 467-481.

Wilson, L. & Bozeman, B. (2004). Market-based management of government laboratories: The evolution of the US National Laboratories' government-owned, contractor-operated management system. *Public Performance & Management Review*, 28(2), 167–185.

Zacharewicz, T., Sanz-Menéndez, L. & Jonkers, K. (2017). *The internationalisation of research and technology organisations*. Seville: Joint Research Centre of the European Commission.

Governmental mobility programmes and academic performance in Spain

Patricia Alonso Álvarez[1], Jorge Mañana Rodríguez[2] and Elías Sanz Casado[2]

[1] *patalons@pa.uc3m.es*
LEMI Research Group, Carlos III University of Madrid, 126 Madrid Str., 28903, Getafe (Spain)

[2] *jmananar@pa.uc3m.es, elias@bib.uc3m.es*
Research Institute for Higher Education and Science (INAECU), Carlos III University of Madrid, 126 Madrid Str., 28903, Getafe (Spain)

Abstract

This research in progress analyses the scientific output of the beneficiaries of one of the most important post-doc mobility programmes in Spain, the José Castillejo call. This research has two main objectives: a) to analyse the volume of international collaboration with the host countries at the medium and long term and b) to test the hypothesis stating that articles written in collaboration with the host countries present greater scientific impact (in terms of citations per document) than those not signed in collaboration. For the period 2009-2017, we retrieved the data on the beneficiaries of the programme and identified their publication set in Web of Science Core Collection (using ORCID codes as input data), obtaining 17,182 registers. A set of regular expressions using the statistical software R allowed us to attribute the output to the beneficiaries with an accuracy rate of 91.58%. The analysis concludes that there are statistically significant differences in citations per document, which favour documents in collaboration. We also find that collaborations persist far beyond the medium term. Several further research issues derived from the initial results are also detailed in this contribution.

Introduction

Scholars' mobility has been largely studied in academic literature (Czaika and Orazbayev, 2018; Kim, 2017; Pheralli, 2011). However, the individual effect of mobility among scholars is yet to be examined. Recent works have evidenced its beneficial effects in terms of scientific performance, including the creation of social and professional networks, improving their scientific production and impact, and foster their prestige (Aksnes et al, 2013; Fanfmeng 2016; Horta, Jung & Santos, 2019, Su, 2011). Still, these results have only been proven for very specific populations and countries. While mobility among Spanish scholars has been addressed in previous literature (De Filippo, Sanz & Gomez, 2009), the impact of government mobility programmes is still understudied. In some countries, talent and government mobility programmes have been proven to have a positive effect on the beneficiaries' careers and research outputs (Liu, Wang & Xu, 2020). This paper (work-in-progress) provides preliminary results of the impact of Spanish governmental mobility programmes in the scientific performance of its beneficiaries, both during the mobility period and afterwards.

For this purpose, we study the scientific production of all the beneficiaries of two public mobility grants: postdoctoral mobility in foreign centres" programme and the "José Castillejo grant for mobility in foreign centres", which together cover the 2009-2017 period. Although the positive effects of these two programmes are manifest for academic career progression in the Spanish system –mobility is one of the criteria of Spanish University System Quality Observatory of the National Agency for Quality Assessment and Accreditation for progressing as an academic or researcher–, the material results of the mobilities have been overlooked. The beneficiaries of the mobility programmes have the opportunity to spend between 3 and 6 months in a foreign University. Prior literature suggests that this opportunity would lead to international collaboration, and higher citation rates and productivity (Aksnes et al, 2013; Su, 2011).

This preliminary study examines the beneficiaries' academic performance during and after the research visit. We use the beneficiaries' production in the Web of Science (WoS) and name disambiguation techniques to attribute each WoS' document to its author. Author disambiguation has been and still is an evolving field in bibliometric studies (Smalheiser & Torvik, 2009; Ferreira, Gonçalves & Laender, 2012; D'Angelo & Van Eck, 2020). The approach of this paper is to classify WoS' documents taking only the affiliation value without using any other information of the metadata such as the author's email, cited references, or other papers' metadata. Inspired by Torvik (2015), we create a name variations dictionary of each beneficiary using automated methods and apply it to the affiliation value of the downloaded WoS' documents.

Preliminary results shown here do not attempt to perform a full analysis of the grants' beneficiaries' performance against no beneficiaries, but to examine patterns among beneficiaries themselves. We find that the percentage of collaboration between the beneficiaries and their countries of destiny is lower than expected, but the impact (measured in citations) of those who have mobility collaborations is indeed higher than the impact of beneficiaries who do not collaborate with their destination country. Production differences have not been found between the two groups.

Data and methods

Data

The data utilized in this analysis were obtained from the web of the Spanish Ministry of Science and Innovation. In particular, the data corresponds to the beneficiaries of the two consecutive mobility programmes: the "postdoctoral mobility in foreign centres" programme, which covers the period 2009-2010; and the "José Castillejo grant for mobility in foreign centres", covering the 2011-2017 period –except for 2013 as the grant was cancelled for that year. Information on the names of the beneficiaries, the resolution year, the start and end date of the mobility period and the country of destination were extracted in DF(Data File)1.In total, grants were given to 1763 beneficiaries in the period 2009-2017.

The ORCIDs were collected for each beneficiary through a semi-automatic procedure and inputted in the Web of Science. Only articles available at the WoS' Core Collection SSCI, SCI and A&HCI sub-databases were retrieved. The results' metadata on WoS's unique identifier of the article, affiliation, publication date and year were downloaded for a total of 17182 documents and saved as DF2.

Pre-processing and regular expressions

Beneficiaries' names from resolutions and affiliation values from the WoS' metadata were converted to ASCII. The names' string was converted to title case and variables were generated for the complete, first and second name, first and second surname, and the initial of the first name. These variables were added to DF.1.

The names of the beneficiaries were identified from the affiliation value of the downloaded WoS' metadata. For this purpose, the names of the authors of each paper were extracted from the affiliations' strings. An intermediate table was created for the classification in which each document was represented in a different row, and the authors of each article were separated in individual columns. This table was called DF2.b.

Regular expressions were constructed from authors' name variants. In total, we used seven regular expressions and they were applied in the following order, which has been designed to minimize the error:

1. Surname1.*Name_complete.*
2. Surname1.*Name1.*
3. Surname1.*Name2.*
4. Surname1.*Surname2.*
5. Surname1.*Name1_initial.*
6. Surname2.*Name1.*
7. Surname2.*Name2

Affiliation extraction

The regex were applied to DF2.b so the row and the column of the match was identified for each document. The full process was as follows:

1. One loop was generated for each year of the grant to minimize errors and mismatching.
2. For each loop, all regex were applied in the order described above.
3. Each matched regex was coded in DF2.b. For articles that only had one match –more than one match could be generated by co-authorship, but more usually by very common names and surnames in articles with many authors–, two columns were generated: one with the beneficiary's name as in DF.1, and one with the regex that obtained the match.
4. No-matched documents or documents with more than one match were passed to the next regex and looped again.
5. Finally, the loop generated one database with all matched documents (DF3).

Figure I shows the percentage of documents that were classified with this process by year. The larger numbers for the last years might be due errors in the ORCID search procedure. Overall, ORCID codes were estimated to match the subjects in 92% of the cases (using a 95% confidence level and a confidence interval of 10). With the regex method, we estimated an accuracy rate of 91.58% in the classification with the same parameters used for the estimation of ORCID search procedure.

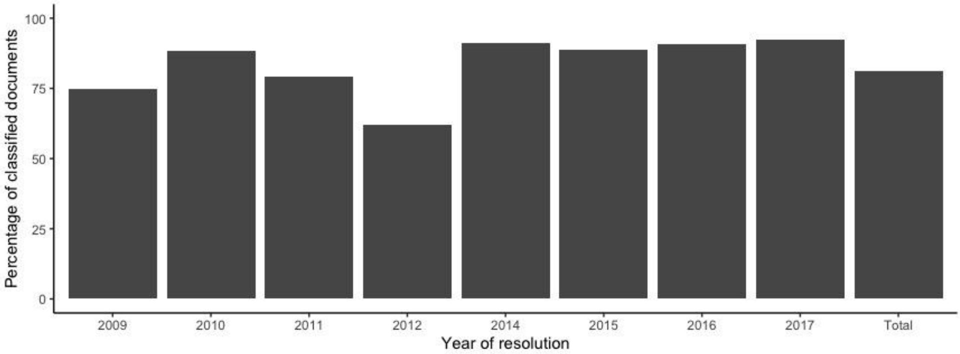


Figure I. Percentage of classified WoS' documents by year and in total.

Identification of production associated with the mobility programme.

The production associated with the mobility programme of each beneficiary was identified using a geographical and a temporal criterion. First, we identified the articles that were written in collaboration with researchers from their destination country using a country dictionary on the affiliations' column of DF3. An important assumption of this step is that articles written in collaboration with researchers from the destination country are likely to have been written during the mobility period or as a result of the collaboration networks developed then.

Secondly, we used the temporal criteria developed in Björk and Solomon (2013) to identify the articles that are more likely to be written during the mobility period considering publishing delay times. Björk and Solomon conclude that the mean period between an article is sent to a journal and published are twelve months (with a standard deviation of 7 months). We calculate two periods. The first approach considers all articles published five months after the start of the mobility in collaboration with researchers from the destination country to evaluate the evolution of the collaborations and the continuity of the networks created. The second method includes all collaboration articles identified before that have been published between five months after the start of the mobility and 19 months after the end of the mobility. This more restricted approach intends to determine the production strictly written during the mobility period. These two approaches will be referred to as Björk and full period respectively.

Results

As can be observed in Figure II, there is an observable decrease in the average percentage of documents in collaboration for the full period. While considering only documents published in the medium term, the collaboration is more stable across concession years.

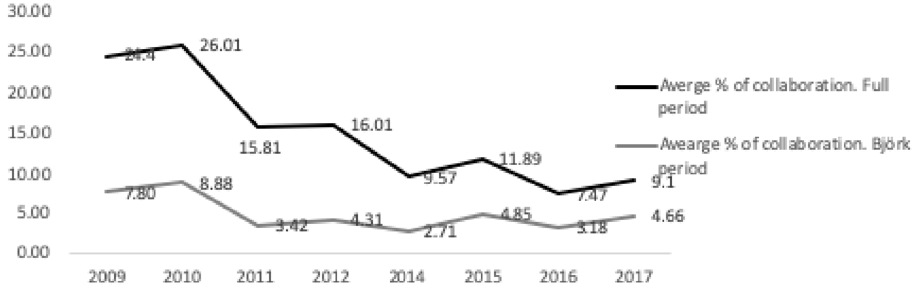


Figure II. Average percentages of collaboration. Full period and medium term

By country, there is an observable shift in the percentages of articles in collaboration. The researchers who made their stays in France, the US and the UK present a strong decrease in the overall percentage of articles in collaboration. Stays in Switzerland, Australia and Portugal are associated, by the opposite, with a growth in the percentages of articles signed in collaboration.

Table I. Compound Annual Growth Rate for the percentages of collaboration (2009-2019 Year of awarded grant).

Country	*N. Docs.*	*Compound Annual Growth Rate (Collaboration %)*
France	1106	-25,69
USA	3954	-23,18
United Kingdom	2767	-12,84
Germany	686	-11,34
Canada	377	-10,63
Netherlands	512	-4,5
Italy	738	-1,01
Switzerland	441	3,26
Australia	564	8,94
Portugal	662	9,58

Removing the documents signed by the researchers when they are affiliated to the stay organization and using values per year, the documents in collaboration (with their host country) present a greater volume of citations than the articles not signed in collaboration. These differences are statistically significant both in the medium and long term (Björk period and full period). The Kolmogorov-Smirnov test shows that the distribution of citations per document are normally distributed across years (Björk period, collaboration [Z=.614; p=.845] Björk period, not in collaboration [Z=.696; p=.717]; full period, collaboration [Z=.653; p=787], full period, not in collaboration [Z=.407; p=.996]). Given the number of authors is very similar each year, we used paired T-test for contrasting the hypothesis that the citations per document are the same for the two groups (articles in collaboration and not in collaboration with the stay country). Citations per document in the Björk period were significantly greater for the articles signed in collaboration (M=40.48; SD=28.37) than for the articles not signed in collaboration (M=19.58; SD=9.58), p=0.025. Citations per document in the full period were significantly greater for the articles signed in collaboration (M=38.96; SD=21.77) than for the articles not signed in collaboration (M=17.45; SD=9.3), p=0.003.

Table II. Citations per document by year. Björk period.

Publication year (Björk period)	*Citations per document collaboration*	*Citations per document NOT in collaboration*
2012	61.73	29.25
2013	52.31	29.05
2014	51.77	27.68
2015	91.01	25.93
2016	24.19	17.05
2017	19.23	12.64
2018	12.66	9.58
2019	10.94	5.46

Table III. Citations per document by year. Full period.

Publication year (full period)	*Citations per document collaboration*	*Citations per document NOT in collaboration*
2012	61.73	29.25
2013	52.37	28.24
2014	56.53	23.21
2015	62.77	20.89
2016	31.81	14.2
2017	20.74	11.27
2018	14.34	8.46
2019	11.44	4.15

Conclusions and discussion

The analysis shows the average of collaborations for the Björk period is low but stable, which means there have been surprisingly few collaborations between the beneficiaries and researchers form the destination countries in all grant resolutions for the this more restrictive period. However, collaborations for the full period show a very different pattern. The higher numbers for the first resolution years show that a considerable number of researchers participating in the research stays show long-term research collaboration with researchers from their host countries.

The documents signed in collaboration, both in the medium and long terms present significantly more scientific impact, measured in citations per document, than the documents not signed in collaboration. Assuming that such collaborations are the one of the stay's outcomes, it can be inferred that the programme is somewhat successful in terms of increasing the impact of the beneficiaries, at least when they engage with scholars from the countries they are visiting.

Despite the clear evidence showing that articles in collaboration show greater citation values, the percentage of articles in collaboration present a steep, linear decrease by award year.

The data analyzed in this study does not allow to conclude any plausible hypothesis but the phenomenon ought to be further studied, since it diminishes the effectiveness of the programme in terms of both collaboration and impact. The slow but noticeable decrease in the percentages of collaboration of the researchers choosing a research organization in France, the UK or the US might be related to a shift in research areas, but this phenomenon together with the increase in the percentage of collaborations in the case of Switzerland, Australia and Portugal might require further research for understanding its causes.

Next steps of this study will also benefit from comparing the results presented here with the analysis of academic performance indicators of researchers that have not been awarded with a mobility grant to address the relative impact of governmental mobility programmes in Spain.

Acknowledgements

This research has been partially funded by the Spanish State R&D Program Oriented to the Challenges of the Society "The Role of Open Science in Spanish Universities: Institutional Transformation and Anticipatory Governance" (ROSSUE, PID2019-104052RB-C21) and the regional Grants for carrying out R&D activity programs between research groups of the Community of Madrid in Social Sciences and Humanities (2019) through the project “Towards the consolidation of inclusive cities, a challenge for Madrid” (CM-INCLUSIVA-CM, H2019/HUM5744).

References

Aksnes, D. W., Rørstad, K., Piro, F. N., & Sivertsen, G. (2013). Are mobile researchers more productive and cited than non-mobile researchers? A large-scale study of Norwegian scientists. *Research Evaluation*, *22*(4), 215-223.

Czaika, M., & Orazbayev, S. (2018). The globalisation of scientific mobility, 1970–2014. *Applied Geography*, *96*, 1-10.

D’Angelo, C. A., & van Eck, N. J. (2020). Collecting large-scale publication data at the level of individual researchers: a practical proposal for author name disambiguation. *Scientometrics*, 1-25.

De Filippo, D., Casado, E. S., & Gómez, I. (2009). Quantitative and qualitative approaches to the study of mobility and scientific performance: a case study of a Spanish university. *Research evaluation*, *18*(3), 191-200.

Fangmeng, T. (2016). Brain circulation, diaspora and scientific progress: A study of the international migration of Chinese scientists, 1998–2006. *Asian and Pacific migration journal*, *25*(3), 296-319.

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. (2012). A brief survey of automatic methods for author name disambiguation. *Acm Sigmod Record*, *41*(2), 15-26.

Horta, H., Jung, J., & Santos, J. M. (2019). Mobility and research performance of academics in city-based higher education systems. *Higher Education Policy*, 1-22.

Kim, T. (2017). Academic mobility, transnational identity capital, and stratification under conditions of academic capitalism. *Higher Education*, *73*(6), 981-997.

Liu, J., Wang, R., & Xu, S. (2020). What academic mobility configurations contribute to high performance: an fsQCA analysis of CSC-funded visiting scholars. *Scientometrics*, 1-22.

Pherali, T. J. (2012). Academic mobility, language, and cultural capital: The experience of transnational academics in British higher education institutions. *Journal of studies in international education*, *16*(4), 313-333.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual review of information science and technology*, *43*(1), 1-43.

Su, X. (2011). Postdoctoral training, departmental prestige and scientists’ research productivity. *The Journal of Technology Transfer*, *36*(3), 275-291.

Torvik, V. I. (2015, November). MapAffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide. In *D-Lib magazine: the magazine of the Digital Library Forum* (Vol. 21, No. 11-12). NIH Public Access.

A Visual Analytics Environment for the Assessment of Information Quality of Performance Models

Marco Angelini[1], Cinzia Daraio[1] and Luca Urban[2]

[1] *{angelini, daraio}@diag.uniroma1.it*

DIAG Department, Sapienza University of Rome, Via Ariosto, 25 00185, Rome (Italy)

[2] *urban.1651869@studenti.uniroma1.it*

DIAG Department, Sapienza University of Rome, Via Ariosto, 25 00185, Rome (Italy)

Abstract

In this paper we extend the flexibility of a visual analytic approach featured to performance model's development to include data quality procedures and tests. The resulting environment is helpful to guide the user to an Information Quality-aware development of Performance models. This interactive visual analytics environment offers to the user the possibility to produce and compare information quality-aware indicators.

1. Introduction and background

In the last decade, the rapid increase in the production, communication and evaluation of research have been signs of a transformation. This transformation has changed also the way we conceive and measure performance indicators. Generally speaking, performance is the result (outcome) obtained by an activity and indicators are results of mathematical operations with data. When indicators are used in an evaluation process, they are called metrics. The definition and measurement of performance indicators require the formulation of a performance model., i.e., the description of the conceptual, methodological and data dimensions capable to assess the outcome of given activities of the units under analysis, with respect to some predefined goals. An example of performance model can be an efficient frontier model that describes how well universities produce their scientific and teaching outputs with respect to inputs and other factors (i.e., including staff, public and private funding, and so on).

Despite the various innovations introduced with big data, machine learning and altmetrics, the role of the *user* of metrics, and her interactions in the development and evaluation phase of performance models have not received the great attention they deserve. In addition, important aspects for the usability of data and information, such as the different dimensions of *data and information quality*, have frequently been overlooked, making the developed performance measurement systems rigid, fragile and inconsistent.

Koltay (2016), discussing data quality and research data management, shows that data governance and data literacy are two important constituents to keep into account when developing a knowledge base. “Applying data governance to research data management processes and data literacy education helps in delineating decision domains and defining accountability for decision making. Adopting data governance is advantageous, because it is a service based on standardized, repeatable processes and is designed to enable the transparency of data-related processes and cost reduction. It is also useful, because it refers to rules, policies, standards; decision rights; accountabilities and methods of enforcement. Therefore, although it received more attention in corporate settings and some of the skills related to it are already possessed by librarians, knowledge on data governance is foundational for research data services, especially as it appears on all levels of research data services, and is applicable to big data (Koltay, 2016, p. 303).” Soylu et al. (2017, 2018) describe the added value of visual methods combined with ontology-based data management for query formulation and for making querying independent of users' technical skills and the knowledge of the underlying textual query language and the structure of data. Daraio et al. (2016) show the benefits of an ontology-based data integration approach for data quality in an open environment. Angelini et al. (2020) present the advantages of Visual Analytics for the development of performance models. In this paper we make a step further and extend the flexibility of a visual analytic approach featured to performance model's development to include data quality procedures and tests.

2. Related work

This paper is an extension of a previously published work by the authors (Angelini et al. 2020) that focused on supporting the development of Performance Models through Visual Analytics. Analyzing and steering Data Quality in analytical processes are very relevant activities in computer science and data analysis, where many automatic tools exist that support, to different degree, those tasks. A recent work by Ehrlinger at al. (2019) surveys 667 software tools dedicated to data quality. Among other considerations, the authors report on the large heterogeneity of the tools, with more than half of them working on proprietary solutions, and most of them lacking implementation of important Data Quality dimensions identified in the literature, or not supporting comparability due to the way in which metrics are defined. Interestingly, the authors report a very low presence of visual support for data quality analysis, with existing works focusing more on usability criteria than on supporting the analysis process. Our proposal goes into this direction, to fill this existing gap in the literature.

Few previous works exist that have explored the use of Visualization and Visual Analytics to conduct Data Quality analysis, in particular for supporting the evaluation of research activities. One of the first works is Rajmonda et al. (2005). They present a tool for visual representation of Data Quality called DaVis. DaVis uses a tabular visual representation to show a dataset, highlights inaccuracies and invalid data, and shows difference between versions of a dataset. However, it focuses only on the visual representation and not on providing further analyses lead by visual interactions. Kandel et al. (2012) propose Profiler, a visual analysis tool for assessing data quality issues. Profiler applies data mining methods to flag problematic data and assess the data in context by automatically suggesting a multiple coordinated visualizations environment. The tool is targeted only at tabular data and it still focuses on providing the best visual evidence of problems in the quality of data.

Moving to Visual Analytics approaches, Liu et al. (2018) offer a literature review on Visual Analytics for Data Quality activities, and a framework for conducting data cleansing on four data types (multimedia, text, trajectories and graphs). Gschwandtner et al. (2014) propose a solution for data cleansing of time-oriented data, providing semi-automatic quality checks, visualizations, and directly editable data tables. However, the authors specifically target time-series and do not target the evaluation of research activities. Finally, Bors et al. (2018) present Metrics-Doc, a visual analytic solution for assessing Data Quality that provides customizable, reusable quality metrics in combination with immediate visual feedback. This solution allows for defining specific quality metrics directly into the system (using OpenRefine syntax) and test and identify data quality violations and distribution.

We highlight, as a differentiating point with respect to these approaches, that our proposal is expressively aimed to the evaluation of educational and research activities, considering their specific indicators and semantics that govern this domain. At the same time, our approach shares similar goal in inserting Data Quality features for supporting performance models development.

Daraio et al. (2020) developed a Data Quality approach featured on higher educational data, that may be integrated with research data and other heterogeneous sources through *Sapientia*, the Ontology of Multidimensional Research Assessment. This is what we are implementing in ongoing research.

3. Aim and contribution

This paper exploits Visual Analytics, “the science of analytical reasoning facilitated by interactive visual interfaces” (Cook and Thomas, 2005), focusing on the Data Quality analysis of the measures, indicators, scores that will be used by the analyst as a base for creating and developing a performance model. In this respect, this phase is very important, given the heterogeneity of data sources, the different formats that can still convey similar semantic, and the importance that features selection can have on the definition of a performance model.
Angelini et al. (2020) provide a workflow for a dynamic creating and assessment of a performance model for evaluating research activities. This workflow is based on an ontological modelling of the data sources, instantiated in the *Sapientia* ontology, that align semantically the contents coming from different sources (e.g., Eter, Scopus, Wos and so on). From this step, the Visual Analytics Environment built allows the data exploration and builds on top of it several performance models (e.g., Efficiency models, input/output models) that can be compared and assessed in order to be validated. Figure 1 shows the mentioned functionalities as steps 1 and 3.

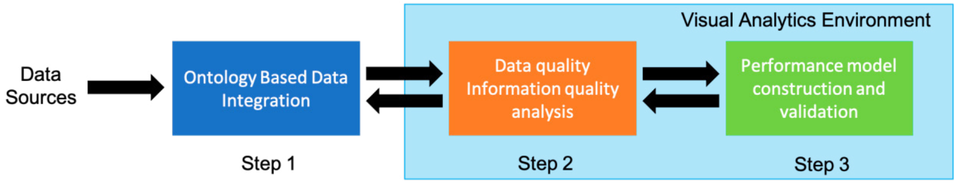

Visual Analytics Environment
Data Sources
Ontology Based Data Integration
Data quality Information quality analysis
Performance model construction and validation
Step 1
Step 2
Step 3

Figure 1: Performance models development workflow

The proposal in this paper leverages on this workflow inserting a new intermediate step (see Figure 1, step 2) that implements the evaluation of the quality of data ingested in the system, during the initial data exploration and/or once the analyst selected a pool of features on which construct the desired models (hence the bi-directional arrows for both model construction and data ingestion). For quality we mean both syntactic properties of the data, like the presence of null or incomplete values or the type of data at hand (e.g., categorical, numerical), and semantic properties, like the fairness of specific features (consistency) and their timeliness.
This intermediate step can reinforce the resulting quality of the developed performance models and can facilitate the check of the fairness of the model or controlling the reliability of the resulting rankings with respect to the statistical significance of the supporting data.
Given the specificity of this additional Data Quality task, the resulting Visual Analytics Environment has been expanded with a tailored solution dedicated to this analysis, split into two main parts: an environment dedicated to provide an overview and exploration of classic Data Quality metrics (Data Quality Environment, DQE), and another environment dedicated to Information Quality evaluation, which refers to how the specific feature can support the creation of a performance model and how a feature contribute to it (Information Quality Environment, IQE). Referring to the DQE, it is constructed with both visual paradigms that are familiar to data quality experts (e.g., Pareto charts) and more abstract representations, like matrix-based visualization or radar-charts. A Data Quality overview from the current version of the DQE is visible in Figure 2: this representation, inspired by the work of Angelini et al. (2018), allows a statistical description of the behavior of the numerical data dimensions with respect to a set of Data Quality metrics using a matrix-like representation.
This interactive visualization represents data dimensions on columns and quality metrics on rows. It visually highlights good quality scores (deep blue color) or bad quality scores (white color) for data dimensions on all quality metrics. This results in the creation of an overview on the overall quality of the dataset through a compact visualization that allows exploration and detailed analysis of subset of pairs *quality metric(s)-data dimension(s)*. The area of an element encodes the variability in scores computed for all the data tuples for each pair *quality metric-*

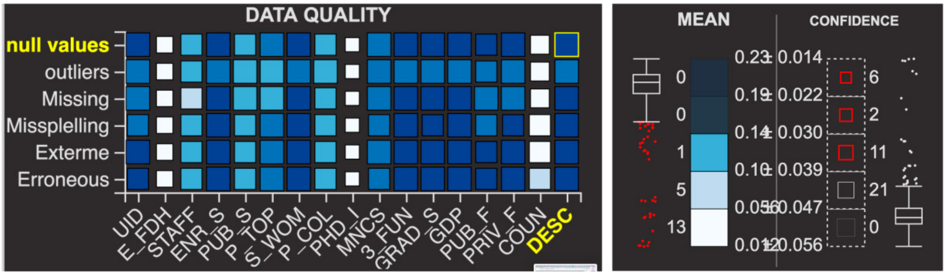


Figure 2: Data quality overview (*left*) with interactive instances exploration panel (*right*)

data dimension. A large area means a narrow confidence interval (stable result with low variability). Conversely, a small area identifies a large confidence interval (high variability).

4. Application to European Higher Education Institutions data (ETER)

On top of the DQE, we are developing also the Information Quality Environment (IQE). We present below some results of the application on the ETER dataset (https://www.eter-project.com/). This dataset is composed by 619 variables about European universities and their activities. In this example, we select the variables that represent the structure and the dimension of the institutions. From these variables we computed the ratios shown in Table 1.

Table 1. Ratios created from the selected variables

Index	Formula	Description
R_1	R&D_EXP / TOT_EXP	Index of funding for Research & Development
R_2	STUD_ENR_I8 / TOT_STUD_ENR	Index of doctorate enrolment
R_3	TOT_EXP / TOT_REV	Index of investments
R_4	COOP_RES / TOT_RES	Index of collaborative research
R_5	GRAD_ISCED8 / STUD_ENR_I8	Index of doctorate graduation in the institution
R_6	(TOT_INC_STUD + INC_STAFF) / (TOT_OUT_STUD + OUT_STAFF)	Index of students and staff mobility
R_7	FOR_ACC_STAFF / TOT_ACC_STAFF	Index of academic foreigners
R_8	CORE_BUDG / TOT_REV	Index of the core funding allocated by the government

Each of these ratios is visualized in the IQE as a *Gauge plot*. Focusing on R_1 (see Figure 4 left), the plot represents the distribution of the ratios (the fraction between the expenditures in Research and Development (R&D) over the total amount of expenditures for one year). From this plot we can say that the institutions on average, spend 23.4% of their total expenditures in R&D, while the upper 5% of the institutions spend more than 60% and the lower 5% of the institutions spend a very little part (around 1%) of the total expenditures in R&D. The analysis can proceed on comparing institutions on this ratio and their R&D expenditures through the use of a *Map plot*. Figure 3 shows the average of the R_1 ratios (on the left) and their R&D expenditures (on the right) aggregated at regional level.

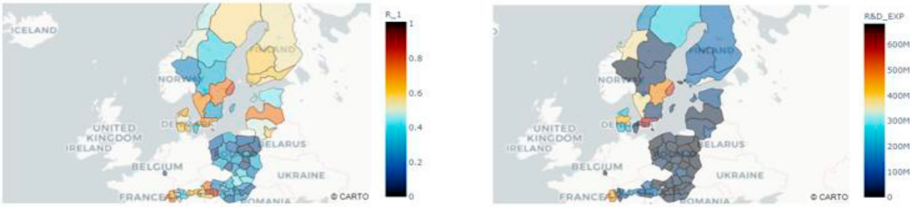


Figure 3: Map plots for R_1 (left) and R&D_EXP (right) variables

We notice that the institutions of Finland, Denmark and Austria are making more efforts (in percentage) in R&D (Fig. 3 on the left), while Polish institutions are expending the less in this field (Fig. 3 on the right).

If we want to compare two variables looking at their correlation, we use the *Heatmap plot* that visualizes the correlation between the target variable (TOT_RES) and all the ratios. See Figure 4 on the right.

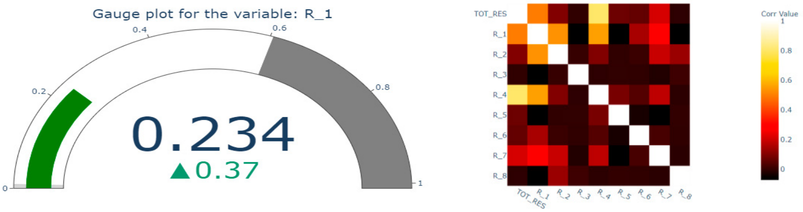


Figure 4: Gauge plot for single variable (left). Heatmap that contains the correlation between the total number of researches and all the ratios (right)

The most correlated ratios with the target variable are R_4 and R_1 ratios, a smaller correlation exists with R_7, while the others are poorly correlated. If we want to see why a variable is highly or poorly correlated to another, we can use the *Scatterplot chart* with the two chosen variables (see Figure 5: TOT_RES variable and R_1 ratios (on the left), and TOT_RES variable and R_2 ratios on the right). By inspecting Figure 5, R_2 results more scattered than R_1.

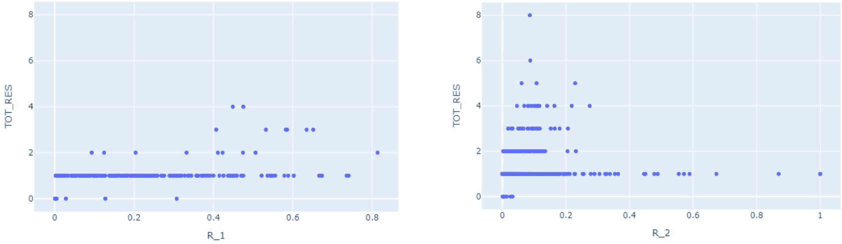


Figure 5: Scatterplots of R_1 and R_2 with respect to the TOT_RES variable

Our IQE includes the *Feature Importance Analysis* (FIA), a multivariate analysis based on the application of *i)* a *Ridge Regression* (with different values of the regularization parameter α) with the chosen features on the target variable, *ii)* Extraction of the *Feature Importance* value, and *iii)* plotting of the *Pareto chart*. An example is shown in Figure 6. While the single variable analysis identified R_1 and R_2 ratios as the best correlated to the TOT_RES variable, the results obtained from the FIA identifies R_4 and R_7 ratios as the most important, with R_4 most prominent (80% of the total importance). Perturbing the regularization parameter, the result does not change much, confirming their validity.

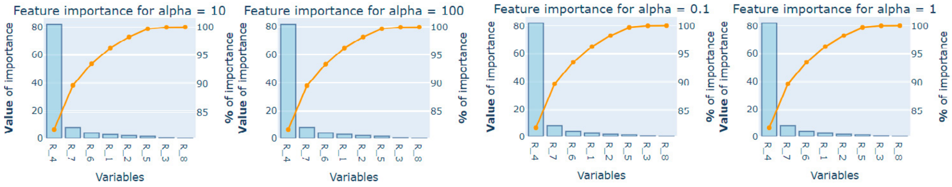


Figure 6: example of interactive Pareto-chart view for feature's importance estimation

5. Concluding remarks

We are working on finalizing the DQE and IQE, integrating them to fully implement the described workflow. It will better support the creation and evaluation phases, providing insights on information and data quality, able to guide the user toward better performance models. A preliminary version of this environment has been used during a Methodological Course on Data

Quality organized within the training activities of the EU RISIS Project (Research Infrastructure for Science and Innovation Policy Studies (https://www.risis2.eu). Further developments and a novel version will be integrated in a web platform for evaluation activities.

Acknowledgments

The funding support of Sapienza Awards n. PH11715C8239C105 and n. RM11916B8853C925 is gratefully acknowledged, together with the RISIS 2, Grant agreement N° 824091, European Union H2020 Project.

References

Angelini, M., Daraio, C., Lenzerini, M., Leotta, F., & Santucci, G. (2020). Performance model's development: a novel approach encompassing ontology-based data access and visual analytics. *Scientometrics*, 125(2), 865-892.

Angelini, M., Fazzini, V., Ferro, N., Santucci, G., & Silvello, G. (2018). CLAIRE: A combinatorial visual analytics system for information retrieval evaluation. *Information Processing & Management*, 54(6), 1077-1100.

Bors, C., Gschwandtner, T., Kriglstein, S., Miksch, S., & Pohl, M. (2018). Visual interactive creation, customization, and analysis of data quality metrics. *Journal of Data and Information Quality* (JDIQ), 10(1), 1-26.

Cook, K. A., & Thomas, J. J. (2005). *Illuminating the path: The research and development agenda for visual analytics* (No. PNNL-SA-45230). Pacific Northwest National Lab. (PNNL), WA, USA.

Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016). The advantages of an Ontology-Based Data Management approach: openness, interoperability and data quality. *Scientometrics*, 108(1), 441-455.

Daraio, C., Bruni, R., Catalano, G., Daraio, A., Matteucci, G., Scannapieco, M., ... & Lepori, B. (2020). A Tailor-made Data Quality Approach for Higher Educational Data. *Journal of Data and Information Science*, 5(3), 129-160.

Ehrlinger, L., Rusz, E., & Wöß, W. (2019). *A survey of data quality measurement and monitoring tools*. arXiv preprint arXiv:1907.08138.

Gschwandtner, T., Aigner, W., Miksch, S., Gärtner, J., Kriglstein, S., Pohl, M., & Suchy, N. (2014). *TimeCleanser: A visual analytics approach for data cleansing of time-oriented data*. In Proceedings of the 14th international conference on knowledge technologies and data-driven business (pp. 1-8).

Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012, May). *Profiler: Integrated statistical analysis and visualization for data quality assessment*. In Proceedings of the International Working Conference on Advanced Visual Interfaces (pp. 547-554).

Koltay, T. (2016). Data governance, data literacy and the management of data quality. IFLA Journal, 42(4), 303-312.

Liu, S., Andrienko, G., Wu, Y., Cao, N., Jiang, L., Shi, C., ... & Hong, S. (2018). Steering data quality with visual analytics: The complexity challenge. *Visual Informatics*, 2(4), 191-197.

OpenRefine (2021), *A free, open source, powerful tool for working with messy data*. https://openrefine.org/ (accessed: January 2021).

Montgomery, D. C. (2020). *Introduction to statistical quality control*. John Wiley & Sons.

Sulo, R., Eick, S., & Grossman, R. (2005). *DaVis: a tool for visualizing data quality*. Posters Compendium of InfoVis, 2005, 45-46.

Soylu, A., Giese, M., Jimenez-Ruiz, E., Kharlamov, E., Zheleznyakov, D., & Horrocks, I. (2017). Ontology-based end-user visual query formulation: Why, what, who, how, and which? *Universal Access in the Information Society*, 16(2), 435-467.

Soylu, A., Kharlamov, E., Zheleznyakov, D., Jimenez-Ruiz, E., Giese, M., Skjæveland, M. G., ... & Horrocks, I. (2018). OptiqueVQS: a visual query system over ontologies for industry. *Semantic Web*, 9(5), 627-660.

Universities and patent trolls: an Empirical Study of University Patent Infringement Litigation in the United States

Grazia Sveva Ascione[1], Laura Ciucci[2], Claudio Detotto[3] and Valerio Sterzi[4]

[1] *graziasveva.ascione@unito.it*
Department of Economics and Statistics, University of Turin

[2] *lciucci@univ-corse.fr*
GREThA UMR CNRS 5113, University of Bordeaux; LISA UMR CNRS 6240, University of Corsica

[3] *detotto_c@univ-corse.fr*
LISA UMR CNRS 6240, University of Corsica; CRENoS

[4] *valerio.sterzi@u-bordeaux.fr*
GREThA UMR CNRS 5113, University of Bordeaux

Abstract

Universities underwent a radical change of paradigm in the last forty years. Since Bayh-Dole Act in 1980, they have been accused of having a growing interest in revenue generation activities such as overzealous patenting, participation in patent auctions and morally doubtful patent enforcement. On the point, some prominent scholars argue that the litigation behavior of universities should be monitored, considering the possible similarity with the so- called patent trolls. Patent trolls are a kind of non-practicing entities (NPEs) whose core business is litigate patents with lower quality almost at the end of their life, trying to maximize their revenues at the expense of the potential plaintiffs. While the harms to innovation made by patent trolls are well-known, the involvement of universities in litigation has not been explored by the literature from an empirically point of view. In this work, we collect data on patents held by universities at the United States and Trademark Office (USPTO) and data on infringement lawsuits filed by universities in the years 1990-2019 to study the characteristics of the litigated patents. We find that universities litigate their patents only sporadically (less than 0.4% of their patents have been used in infringement proceedings) and when the patents are particularly valuable. Moreover, we analyse the fields in which universities litigate, considering that trolls mainly ravage in the ICT sector. Our conclusion supports the idea that universities participating in litigation is a growing phenomenon which should be monitored, but their current behavior does not reflect the strategies of the *litigation* NPEs. Further research is needed to consider the whole universities' portfolio to assess which patents they choose to litigate and the evolution of their strategies over time.

Introduction

University, traditionally been perceived as a relevant actor in the innovation ecosystem, gradually changed its role, integrating its teaching and research mission with the commercial assumptions underlying IP law (Etzkowitz, 1990; Ghosh, 2016). Thus, understanding and analyzing the changes which have taken place in the relationship between higher education institutions (HEIs) and intellectual property is relevant to the whole innovation system (Adams, 1990).

In US, universities have been involved in technology transfer activities since the nineteenth century (Meyer-Thurow, 1982), with an upsurge in patenting activities after the adoption of the Bayh-Dole Act; the increase was also sustained by the increasing importance of software and biotechnology, which were emerging sectors in the 1980s, together with the shifting of norms in academia (Firpo and Mireles, 2018; Rai and Eisenberg, 2003). After Bayh-Dole, universities started establishing TTOs to help researchers with patent applications and to manage licensing revenues.

In an attempt to increase revenues generated by licensing and patenting activities, in recent years some universities have started to pursue "overzealous" strategies to protect their existing patents, by enforcing them in courts, selling them to the highest bidder and making arrangements with patent assertion entities (PAEs) (Fusco et al., 2019; Love et al., 2020). As pointed out by the former Harvard president Deren Bok, the "lure of marketplace" might modify and hamper universities' core values (Bok, 2009), giving incentives to universities to patent inventions which might have or not a commercial value but which for sure could be enforced in courts for their "litigation" value (Frye and Ryan Jr, 2020).

Despite the importance of the topic, very few studies on university patent litigation exist. To our knowledge, only Rooksby (2011, 2012) and Firpo and Mireles (2018, 2020) collect and examine data on patent litigations initiated by universities in the US. In particular, Firpo and Mireles (2018) analyze litigation cases filed between 2000 and 2015 by universities, foundations and non-profit organizations, for a total of 585 cases and had the merit of being the first to explore evidence of strategic (troll-like) behavior in university patent litigation. They find that, contrary to patent trolls, universities assert on average relatively young patents (7.5 years old at the time of litigation); however, they also find that there are numerous patent litigated by universities that are in the final years of their term, suggesting that in some circumstances some universities also assert their patent against a technology that has been already commercialized.

In this research, we use a novel dataset which further investigates the participation of universities in litigation in the US. In particular, we collect patent infringement lawsuits involving universities up to 2019 from Darts-IP and US granted (utility) patents at the USPTO where universities appear as assignees, from Patent Views (2018) and the Patent Assignment Database (version 2017). Our results provide evidence of the increasing participation of universities in litigation activities during the last twenty years. Moreover, our econometric analysis shows that universities litigate the most valuable patents of their portfolio and that are not as much as broad as one could expect from PAEs (Fisher and Henkel, 2012). However, the litigation and patenting behaviour of universities should be closely monitored, by investigating differences across technologies (we observe a significant share of patents litigated in the ICT field), and studying the emergence of new universities (included non-US universities) and investigating the change in their monetisation strategies over time.

The rest of the paper is organized as follows. In the next section we discuss the patent commercialization strategies adopted by universities, with particular attention to their "non-orthodox" monetization behavior and we review the present literature about their participation in litigation. Then, we describe the patent and litigation data and facilitate an empirical understanding of university participation in patent infringement litigation in the US in the years 1990-2019. Afterwards, we run an econometric model, in order to analyze the characteristics of patents litigated by universities in the US. Finally, a discussion of the results and acknowledgment of limitation concludes.

University patents commercialization

Prominent scholars extensively studied the surge for patenting which followed Bayh-Dole, looking for the effects on innovation production and on universities as actors for the first time involved in the challenges of commercializing and licensing their inventions (Mowery and Ziedonis, 2002; Mowery et al., 2002).

Bayh-Dole Act has been harshly criticized by some, being defined as a "one size fits all" law which do not consider that patents from diverse field have different necessities and opportunities for commercialization (Mireles and Firpo, 2018). For instance, patents may play a central role in industries such as pharmaceuticals, while in others, such as IT and software, patents might be not as important or even harmful (Burk and Lemley, 2002). Moreover, licenses do not seem so fruitful for universities either, not enough to justify the current scope of Bayh-Dole. Powers and Campbell (2009) demonstrated that 35% of universities in their sample never realized profitability no matter their investment. In addition, a 2013 Brooking Institution extensive study proved that, in 2012, 84% of universities failed to break even in technology transfer, considering the personnel and filing costs (Valdivia, 2013).

The issue of whether society benefits most if universities patent (or not) their inventions is entangled with another debate about whether universities should try to maximize the revenue stream from their patents (Lemley and Feldman, 2020). Jennifer Washburn (2008) defined the benefits of commercialization as "vastly overblown", suggesting that the increasingly aggressive attitude of universities towards IP appropriation not only would dissuade broad public access, but might also encourage monetization as a mission of universities, thus further disincentivizing dissemination of technologies and thwarting rather than facilitating commercial development (Raj, 2004). Moreover, the increasing interest in commercialization would increase universities' interest in acting like "patent seekers, patent enforcers, and patent policy stakeholders" (Eisenberg and Cook-Deegan, 2018). Many university administrators in the US state that revenue generation is a relevant goal of their technology transfer operations, and that this was already the case in the early 2000s (Thursby et al., 2001; Abrams et al., 2009), in part in response to the increasing threat of dwindling public funding (Firpo and Mireles, 2018, 2020). Recent evidence shows that there has been a significant increase in the monetization activity around the related patents, especially in the US, and that university TTOs appear to be "revenue-driven with a single-minded focus on generating licensing income" (Kesan, 2008). University pursuit of licensing to increase revenues generated consequences which go beyond the greater amount of resources invested in TTOs, the orientation towards applied research and exclusive licenses. For instance, looking for even greater revenues, universities have even been tried to organize patent auctions, selling even to PAEs (Ledford, 2013, Cahoy et al., 2016).

Universities and patent trolls: an ambiguous relationship

PAEs are owners of patents who file or acquire patents without having any interest in actually developing the product. The activity of PAEs has been largely analyzed by scholars in terms of patents' acquisitions and litigation, with a special eye on the latter, considering its potentially harmful impact on innovation (Bessen and Meurer, 2013; Chien, 2013; Kiebzak et al., 2016). Allison et al. (2010) showed that almost 50% of the most litigated patents are owned by PAEs, making them the most litigious actor on the market and pointing out the importance of monitoring their behavior (Lemley and Melamed, 2013).

In general, PAEs are accused to focus their business mostly on litigating patents related to product already developed, harming innovation and damaging companies and individuals which do not have enough resources to fight them in court: these are called patent trolls (Chien, 2012; Feldman and Lemley, 2015; Cohen et al., 2019). With the threat of expensive litigation process, patent trolls convince inventors to pay them a license to use the litigated patent, regardless of the supposed infringement being true or not. Thus, trolls extract high costs in the form of impeding innovation and pushing up costs for consumers for goods and services covered by patents (Firpo and Mireles, 2018). At the same time, PAEs are also described as "middleman", bringing technologies from inventors to people who can implement them (Fisher and Henkel, 2012; Steensma et al., 2012) and helping small inventor to monetize their IP assets (Haber and Werfel, 2016); hence, they play a relevant administrative role, identifying relevant technology and collecting royalties, complementing in that way the scarce expertise of some patent holders in the latter activity.

The US Federal Trade Commission report on PAEs (FTC, 2016) proposes to distinguish two kind of entities under the umbrella of PAEs. *Portfolio PAEs* are large patent aggregators that usually hold a great amount of patents which are not necessarily used for litigation purposes; their goal is to promote technology transfer, acting as middleman between demand and supply of knowledge. *Litigious PAEs* are smaller PAEs set up as limited liability companies (LLCs) which sue potential licensees and usually settle quickly after entering into negotiations with them. The litigious PAEs have been often accused of starting litigation with the sole purpose of extracting rent, managing to be profitable even if they assert low quality patents, and are labelled like “patent troll”.

Universities have been related to patent trolls for different reasons. First of all, as patent trolls, universities are *non-practicing entities* which have been accused to sell or license their patents without the interest in transferring the underlying technology (Cordova and Feldman, 2015; Feldman and Ewing, 2012; Love et al., 2020); secondly, they have been recently blamed of acting as trolls, (Lemley, 2016; Firpo and Mireles, 2018), being involved into patent litigation with the sole aim of increasing their revenues.

With respect to the first criticism, recent evidence shows that the market of university inventions, by analyzing US patent transfers where universities appear as sellers, between 2012 and 2017, Love et al. (2020) find tht only 10 percent of them appear to transfer knowhow in addition to the patent assets listed therein. Similarly, XXX shows that Intellectual Ventures (IV), one of the biggest patent holding company in US and notorious PAE, opened up about its relationship with more than 400 universities (whose 60 American), but only two of them eventually led to commercial products.

With respect to the second criticism, it is now evident that universities are involved in the patent ecosystem not only as patent owners and licensors, but also as litigants, chasing lavish monetary rewards. In particular, universities have been accused of being patent trolls also because they can enjoy some of their legal benefits: they do not incur the risk of being counterattacked (Lemley, 2007). Beyond seeking patents not always useful for commercialization, anedoctal evidence shows that universities started to enforce their patents in litigation against firms that have already developed successful commercial products (just as patent trolls would do). Many of these lawsuits have led to very large settlement or damages awards. One such case involves Boston University. Boston University sued more than thirty defendants, including Apple, for infringing a late term patent covering LED technology.

Critics also underline the inconsistency between universities values and the troll-like attitude: universities live on tax subsides, so they are supposed to pursue public good through conversation, research and teaching, excluding morally doubtful activities such as patent litigation. In addition, the money invested in patent litigation might harm universities, because litigation would subtract money and time to university personnel. Patent enforcement might decrease productivity of TTOs as well, generating an extra cost for consultancy with attorneys about strategies for ligation; at the same time, they involve TTOs personnel in document production and witnessing at trial, who would otherwise be busy with marketing, search and negotiation activities, thereby reducing the amount of licensing. Hence, patent litigation might have an adverse effect on university licensing activity itself (Shane and Somaya, 2007).

In our paper, we collect data on US patents held by universities and we analyze their characteristics in order to investigate whether the universities' behavior can be compared to that of patent trolls. In particular, we do expect universities behave as patent trolls if we observe that they litigate patents that are of low-quality but at the same time broad, and so more likely to be infringed.

Data sources and key figures

Dataset construction

The dataset used for this research is created from several sources. First, patent infringement actions involving universities are collected. Second, we match these data with other publicly available databases on US patent data (USPTO); to do this, we rely on different patent databases - Patent Views (Version 2018), Patent Assignment Database (version 2017), EEE-PAT - which we use to identify all patents assigned to universities (and their TTOs) and their characteristics. By limiting our investigation to granted patents filed since 1990, our sample contains 156,492 university patents, of which 107,114 include US universities among the applicants. Third, we rely on the US-OECD patent quality database (version 2019), in order to retrieve information on patent characteristics.

Data on litigation cases were collected from Darts-IP, which provides extensive coverage of IP verdicts across all the most important IP markets over the world, and which allowed us to identify the types of actors (company, university, PAE, person, authority). For this study we collected all patent infringement actions in the United States involving universities (including their TTOs and schools of medicine) as plaintiff, up to December 31, 2019 inclusive. Patent litigation data from Darts-IP contain information about where the cases were filed, the relevant dates, the patent number, the names of plaintiffs and defendants, the outcome (if any), as well as other information associated with the cases. After having collected all the patent infringement cases involving universities, we cleaned and harmonized plaintiffs' and defendants' names, taking into account different spellings and to put together each university with its affiliated research entities (for example, the Arizona State University and Arizona Research Foundation). The final database accounts for 574 infringement actions filed in the US and 813 litigated patents granted at the USPTO, of which 521 are university patents and 292 are patents that are assigned to other types of organization classified as co-plaintiffs of universities in the patent infringement lawsuit. The average number of patents disputed per case is 2.76 and about one third of these cases concern more than two patents.

Descriptive statistics

Aggregate statistics show the increasing number of cases of patent litigation by universities. Figure 1 shows the evolution over time of the number of new patent infringement litigation cases involving universities as plaintiffs from 1990 to 2019 (figures for 2019 may be incomplete due to updating delays), where the year corresponds to the year in which litigation cases start: and shows a constant rate change starting from the late 1990s, with an average annual number of new cases between 20 and 40 in the years 2005-2019, and a noteworthy peak in 2013 when we observe more than 90 new cases.

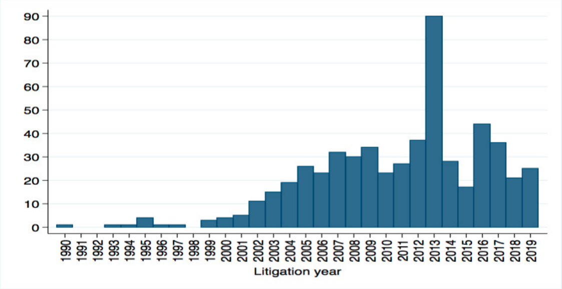


Figure 1. Number of university infringement actions filed in US (1990-2019)

Notes: This distribution is related to all patent infringement actions filed in the US by universities in the period 1990-2019. The litigation year corresponds to the year in which cases start.

Most litigious universities

The 558 identified patent infringement actions involved a total of 93 public and private universities among the plaintiffs; 48 of them were involved in one or two cases, while 15 universities were involved in ten litigations or more (see Table 1). The majority of these are American universities, with the exception of *National Cheng Kung University* (Taiwan), the *University of Strathclyde* (UK) and *Queen's University at Kingston* (Canada). The most active university is *Boston University* with 42 litigation cases, closely followed by the *University of California*, the *Massachusetts Institute of Technology* and the *University of Texas System*, with respectively 37, 32 and 26 cases. Confronting these figures with the ranking of the same universities in terms of patent production, it is possible to observe that, consistently, top performers in patent ranking (such as, University of California, Massachusetts Institute of Technology, University of Texas System etc.) present a higher number of litigation cases, with the exception of Boston University, which is involved in the greatest number of litigations but in proportion has few patents (see Table 1).

Table 1: Top universities, by number of patent infringement suits in United States (1990-2019)

Research institutions	*Country*	*Type*	*No. of cases*	*No. of patents filed*	*Rank pat.*	*Pat. litigated*
Boston University	United States	Private	42	411	55	2
University of California	United States	Public	37	11483	1	49
Massachusetts Institute of Technology	United States	Private	32	4570	3	45
University of Texas System	United States	Public	26	3634	4	16
University of Wisconsin–Madison	United States	Public	23	2611	5	8
University of South Florida	United States	Public	17	935	25	3
National Cheng Kung University	Taiwan	Public	15	463	64	8
Emory University	United States	Private	15	565	44	8
University of Michigan	United States	Public	15	279	98	4
John Hopkins University	United States	Private	15	2016	9	16

Notes: The ranking in the penultimate column is based on the number of patents held by universities at the USPTO since 1990.The last column of the table represents the number of patents (with unique identifier) litigated in the considered time

Technology

Table 2 shows the number of patents held by universities in the US and granted since 1990, and those that have been litigated (0.3%), by technology field according to the 5-sector WIPO classification.

Table 2: Frequencies of US university patents by technology fields

	Litigated patents		*All patents*	
	Nb. (A)	*Percent*	*Nb. (B)*	*Percent*
Electrical Eng.	122	*24.65%*	37,779	*24.03%*
Instruments	118	*23.84%*	35,292	*22.45%*
Chemistry	229	*46.26%*	75,843	*48.24%*
Mechanical Eng.	20	*4.04%*	6,881	*4.38%*
Other fields	6	*1.21%*	1,409	*0.90%*
Total	**495**	***100%***	**157,204**	***100%***

Notes: Only US university patents granted since 1990 are considered and patent infringement actions initiated by universities (as plaintiffs or co-plaintiffs) in the years 1996-201

The largest number of patents filed and litigated by universities in the US is in *Chemistry* (46% of all litigated patents and 48% of university patents), which is also the technological field where they file the largest number of patents (48% of university patents). Interestingly, following the general trend in the patent litigation landscape, while universities litigated only a share of their patents in the ICT sector during the first decade of 2000s, in the last ten years they have increased their presence in this sector (Figure 2).

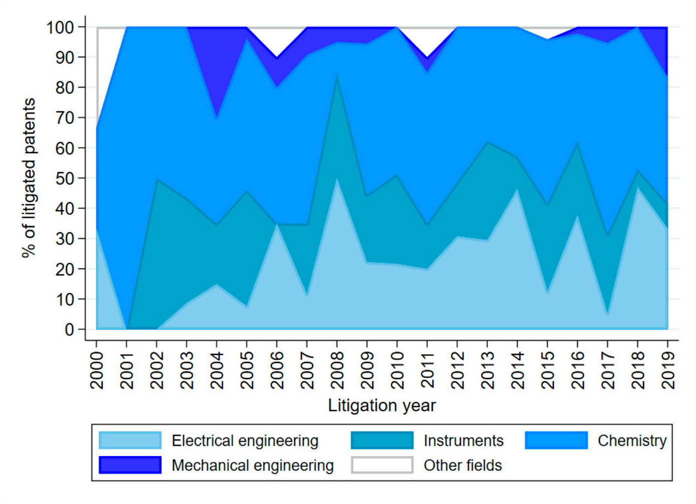


Figure 2. Evolution of the share of litigated patents by sector

Notes. Patent infringement actions initiated by universities (as plaintiff or co-plaintiff) in the period 2000-2019.

Empirical Analysis: litigation decision analysis

In this section we study the litigation strategies of universities in the US by providing a broad-based statistical characterization of patent cases filed in which they appear as plaintiffs. In particular, looking at the main characteristics of the patents under litigation, the aim of this study is to identify the characteristics of litigated patents, with respect to non-litigated patents the universities have in their portfolios.

In particular, we analyse the factors explaining the likelihood of a patent being litigated; in doing so, we consider all university patents filed since 1980 and granted until 2014. We discard from the analysis any patents granted after 2014 because only a few have been litigated later and because they are subjected to a substantial truncation of the number of forward citations which we use as a proxy for technological patent quality. The final dataset contains 134,421 patents, of which 503 have been litigated by universities.

In our econometric exercise, the response variable (*Litigation*) is thus a dichotomous variable whereby litigated patents are designated as "1" while non-litigated ones are coded with "0". Hence a Logit model approach is deemed appropriate here:

$$Prob(Litigation_i = 1) = \mathrm{F}(X_i) = \frac{e^{\beta' X_i}}{1+e^{\beta' X_i}} \qquad (1)$$

In other words, the probability of a litigation is assumed to be a function of a vector of explanatory variables, *X*. Our main variable of interest is the technological importance of the invention (*Fwd_cits*5), which we proxy with the number of citations received by the focal patent in a window of five years starting from the filing date. Citations received by a patent are one indication that an innovation has contributed to the development of subsequent inventions (Henderson et al., 1998). Moreover, we include a set of indicators of patent value (*Claims*, *Bwd_cits*, *Renewals* and *Npl_cits*). In general, the higher the value of these indicators, the higher the expected value of the patent (Squicciarini et al., 2013). *Claims* accounts for the number of claims of a given patent, which define novel features of the invention. Generally, it is associated with the technological breadth or the market value of a patent and reflects a larger risk of conflict with competitors (Tong and Frame, 1994). Similarly, *Renewals* is also a proxy for patent value, indicating the number of years a patent is valid. Since maintaining patent protection over time is costly, we assume that valuable patents pay at least for their own renewal and that the more valuable patents will be renewed for a longer time (Sterzi et al., 2019). *Bwd_cits* looks at backward citations, considering all potential sources of knowledge of a given patent, such as prior patents and scientific works (Harhoff et al., 2003). A low number of backward citations could indicate that the invention is in a relatively new and uncrowded technology area (Harhoff et al., 2003; Harhoff and Reitzig, 2004; Lanjouw and Schankerman, 2004). We thus expect the likelihood of litigation to increase with the number of backward citations. *Npl_cits* measures the size of the non-patented literature cited by the patent, which denotes the technical closeness of an invention to scientific knowledge and is found to be correlated with patent value (Branstetter, 2005).

Patent_scope estimates the breadth of the patent and is measured by counting the number of distinct 4-digit IPC classes the invention is allocated to (Lerner, 1994) and it is thus expected to be correlated with the probability that the patent can be infringed (Lerner, 1994; Fischer and Henkel, 2012).

Next, we control for applicants' country, by including a dummy (*US*) which takes the value of one when at least one university among the applicants comes from the United States. This variable is meant to control for the different costs of trial relative to the cost of settlement for domestic and foreign applicants (universities) and for the disadvantage for foreign universities in detecting infringements in the US market. Moreover, since there are pronounced differences in litigation rates across technology fields and years (Lanjouw and Schankerman, 2001), we also control for technological sector (WIPO 35 technological classes) and grant year fixed effects.

The first column of Table 4 provides the main descriptive statistics of the variables under study. In the overall period observed, litigated patents account for the 0.37% of the whole dataset.

Columns (2)-(5) of Table 4 provide the results of the logistic regression. The different specifications differ little in terms of their explanatory power, where year dummies are included in columns (2) and (3) and technological field dummies are added in columns (2) and (4). Given the values of both pseudo-R^2 and AIC, Column (2) is taken as the reference equation where we control for both annual and technological fixed effects.

Although the overall rate of litigation is low, there is a wide variation across patents in their exposure to litigation risk. First of all, the patent quality is found to be positively correlated with the litigation outcome. All the coefficients of patent value indicators are positive and significant, indicating that a patent's market value increases litigation propensity. On the contrary, the estimated coefficient for *Patent_scope* is non-significant. Overall these results seem to reject the hypothesis that universities behave as patent trolls. As regards the control variables, US patents filed by US universities are more likely to be litigated than patents filed by non-US universities: the odds of being litigated for US university patents is 2.620 times that of non-US university patents.

Table 4: Litigation analysis results. Dependent variable: Litigation (LOGISTIC Model)

	(1)	(2)	(3)	(4)	(5)
VARIABLES	*Mean (sd)*#	*Logit (I)*+	*Logit (II)*+	*Logit (III)*+	*Logit (IV)*+
Litigation	0.003	—	—	—	—
	(0.061)	(—)	(—)	(—)	(—)
US (=1)	0.717	0.964***	0.973***	1.109***	1.102***
	(0.450)	(5.477)	(5.494)	(6.346)	(6.259)
Fwd_cits5	11.223	0.003***	0.003***	0.003***	0.003***
	(34.484)	(10.870)	(9.944)	(11.610)	(10.980)
Patent_scope	2.422	-0.023	-0.058**	-0.037	-0.068***
	(1.557)	(0.886)	(2.268)	(1.368)	(2.622)
Claims	17.961	0.008***	0.008***	0.008***	0.008***
	(14.574)	(4.872)	(5.541)	(4.941)	(5.546)
Bwd_cits	14.620	0.011***	0.012***	0.009***	0.010***
	(21.090)	(8.188)	(9.659)	(6.826)	(8.509)
Npl_cits	16.489	0.013***	0.011***	0.011***	0.009***
	(24.395)	(8.281)	(7.323)	(7.463)	(6.701)
Renewal	10.304	0.086***	0.095***	0.142***	0.141***
	(3.774)	(5.130)	(5.678)	(9.686)	(9.618)
Constant	—	-14.302***	-13.772***	-9.173***	-8.624***
	(—)	(14.261)	(14.080)	(24.870)	(37.231)
Observations	134,421	134,421	134,421	134,421	134,421
Field dummies	—	YES	NO	YES	NO
Year dummies	—	YES	YES	NO	NO
Pseudo R2	—	0.106	0.085	0.095	0.074
AIC	—	6,042.469	6,128.633	6,071.951	6,146.518

Notes: Unit of observation: patent. Grant years: 1980-2014; #Standard deviations in parentheses; +Robust z-statistics (in absolute value) in parentheses; ***p<0.01, **p<0.05, *p<0.1

Discussion and conclusions

Over the past 40 years, there has been a strong growth in university patent filings at the USPTO, due mainly to US institutions increasing their activity in response to the Bayh Dole Act, as well as to the explosion of the software and biotechnology sector and, since the 2000s, to the sharp rise in foreign universities. University patents granted by the USPTO accounted for about 1% of the total number of issued patents before the 1990s and now stand almost at 4%. At the same time, the role of TTOs evolved from ensuring patent protection and simple managing of licensing activities to a more "patent-centric" vision (Kesan, 2008; Pilz, 2020; Frye and Ryan Jr, 2020), in the attempt of achieving sometimes competing objectives (Sherer and Vertinsky, 2020). The extreme consequences of that vision include a tendency to patent as much as possible to relieve the economic pressure to cover their costs, as well as TTOs participating in patent litigation as "buffer" organizations between the asserter university and the defendants (Rooksby, 2011).

Despite the increasing attention devoted to the modern surge in patent filings and to the participation among patent trolls in the patent market, an equal attention has not been devoted to the characteristics and consequences of the increasing interest of universities for litigation activities - except a few cases (Rooksby, 2011; Firpo and Mirales, 2018, 2020).

Our paper goes further in depth on the topic, by collecting data on patents filed by universities at the USPTO and infringement lawsuits involving universities in the years 1990-2019. Our findings can be summarized as follows. First, we show an increasing participation of universities in litigation activities, despite litigating only a share (less than 0.4%) of their patent portfolio. Second, we observe an increasing participation of universities in court in the ICT industry, which is traditionally the sector in which patent trolls litigate the most. Third, universities litigate valuable patents but that are not broader than the non-litigated patents in their portfolios.

Our first evidence seems to corroborate the idea that universities, although sharing some characteristics with trolls, are not trolls. However, this study also presents some important limitations, which constitute avenues for further research. First, the quality of patents litigated by universities should be considered as a factor which could change over time, in order to take a closer look at possible universities' strategy changes. In addition, the same is true for the fields involved in litigation: despite the current prevalence of the *Chemistry* sector, changes over time should be taken into consideration. Second, other litigation strategies (such as the time in which universities start to litigate, the venue and the number of defendants per litigated patent) should also be taken into account.

Bibliography

Abrams, I., Leung, G., & Stevens, A. J. (2009). How are us technology transfer offices tasked and motivated-is it all about the money. *Research Management Review*, *17*(1), 1–34.

Adams, J. D. (1990). Fundamental stocks of knowledge and productivity growth. *Journal of political economy*, *98*(4), 673–702.

Allison, J. R., Lemley, M. A., Moore, K. A., & Trunkey, R. D. (2003). Valuable patents. *Geo. Lj*, *92*, 435.

Allison, J. R., Lemley, M. A., & Schwartz, D. L. (2017). How often do non-practicing entities win patent suits? *Berkeley Technology Law Journal*, *32*(1), 237–310.

Allison, J. R., Lemley, M. A., & Walker, J. (2009). Extreme value or trolls on top-the characteristics of the most-litigated patents. *U. Pa. L. Rev.*, *158*, 1.

Allison, J. R., Lemley, M. A., & Walker, J. (2010). Patent quality and settlement among repeat patent litigants. *Geo. LJ*, *99*, 677.

Bessen, J., & Meurer, M. J. (2013). The direct costs from npe disputes. *Cornell L. Rev.*, *99*, 387.

Bok, D. (2009). *Universities in the marketplace: The commercialization of higher education* (Vol. 49). Princeton University Press.

Branstetter, L. (2005). Exploring the link between academic science and industrial innovation. *Annales d'Economie et de Statistique*, 119–142.

Burk, D. L., & Lemley, M. A. (2002). Is patent law technology-specific. *Berkeley Tech. LJ*, *17*, 1155.

Cahoy, D. R., Kwasnica, A. M., & Lopez, L. A. (2016). The role of auctions in university intellectual property transactions. *Duq. L. Rev.*, *54*, 53.

Chien, C. (2013). Startups and patent trolls. *Stan. Tech. L. Rev.*, *17*, 461.

Cohen, L., Gurun, U. G., & Kominers, S. D. (2019). Patent trolls: Evidence from targeted firms. *Management Science*, *65*(12), 5461–5486.

Cordova, A. K., & Feldman, R. (2015). Universities and patent demands. *Journal of Law and the Biosciences*, *2*(3), 717–721.

Eisenberg, R. S., & Cook-Deegan, R. (2018). Universities: The fallen angels of bayh-dole? *Daedalus*, *147*(4), 76–89.

Etzkowitz, H. (1990). The second academic revolution: The role of the research university in economic development. In *The research system in transition* (pp. 109–124). Springer.

Feldman, R., & Ewing, T. (2012). The giants among us. *Stanford Technology Law Review*, *1*.

Feldman, R., & Lemley, M. A. (2015). Do patent licensing demands mean innovation. *Iowa L. Rev.*, *101*, 137.

Feng, J., & Jaravel, X. (2020). Crafting intellectual property rights: Implications for patent assertion entities, litigation, and innovation. *American Economic Journal: Applied Economics*, *12*(1), 140–81.

Firpo, T., & Mireles, M. S. (2018). Monitoring behavior: Universities, nonprofits, patents, and litigation. *SMUL Rev.*, *71*, 505.

Firpo, T., & Mireles, M. S. (2020). Currents and crosscurrents in litigation of university and nonprofit related patents: is there a coming wave of patent litigation involving those patents? In *Research handbook on intellectual property and technology transfer.* Edward Elgar Publishing.

Fischer, T., & Henkel, J. (2012). Patent trolls on markets for technology–an empirical analysis of npes' patent acquisitions. *Research Policy*, *41*(9), 1519–1533.

Frye, B. L., & Ryan Jr, C. J. (2020). Technology transfer and the public good. In *Research handbook on intellectual property and technology transfer.* Edward Elgar Publishing.

Fusco, S., Lissoni, F., Martinez, C., & Sterzi, V. (2019). Monetization strategies of university patents through paes: an analysis of us patent transfers. *Available at SSRN 3410086*.

Ghosh, S. (2016). Are universities special. *Akron L. Rev.*, *49*, 671.

Haber, S. H., & Werfel, S. H. (2016). Patent trolls as financial intermediaries? Experimental evidence. *Economics Letters*, *149*, 64-66.

Harhoff, D., & Reitzig, M. (2004). Determinants of opposition against epo patent grants—the case of biotechnology and pharmaceuticals. *International journal of industrial organization*, *22*(4), 443–480.

Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research policy*, *32*(8), 1343–1363.

Kesan, J. P. (2008). Transferring innovation. *Fordham L. Rev.*, *77*, 2169.

Kiebzak, S., Rafert, G., & Tucker, C. E. (2016). The effect of patent litigation and patent assertion entities on entrepreneurial activity. *Research Policy*, *45*(1), 218–231.

Lanjouw, J. O., & Schankerman, M. (2001). Characteristics of patent litigation: a window on competition. *RAND journal of economics*, 129–151.

Lanjouw, J. O., & Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal*, *114*(495), 441–465.

Ledford, H. (2013). Universities struggle to make patents pay. *Nature*, *501*(7468), 471–472.

Lee, P. (2013). Patents and the university. *Duke Law Journal*, 1–87.

Leiponen, A., & Delcamp, H. (2019). The anatomy of a troll? patent licensing business models in the light of patent reassignment data. *Research Policy*, *48*(1), 298–311.

Lemley, M. A. (2007). Are universities patent trolls. *Fordham Intell. Prop. Media & Ent. LJ*, *18*, 611.

Lemley, M. A., & Feldman, R. (2020). Is patent enforcement efficient? In *Research handbook on intellectual property and technology transfer.* Edward Elgar Publishing.

Lemley, M. A., & Melamed, A. D. (2013). Missing the forest for the trolls. *Colum. L. Rev.*, *113*, 2117.

Lerner, J. (1994). The importance of patent scope: an empirical analysis. *The RAND Journal of Economics*, 319–333.

Love, B. J., Oliver, E. and Costa, M. (2020). Us patent sales by universities and research institutes,in'Research Handbook on Intellectual Property and Technology Transfer', Edward Elg

Lu, J. APACrefauthors (2012). The myths and facts of patent troll and excessive payment: have nonpracticing entities (npes) been overcompensated? *Business Economics*, *47*, 234–249.

Magliocca, G. N. (2006). Blackberries and barnyards: Patent trolls and the perils of innovation. *Notre Dame L. Rev.*, *82*, 1809.

Meyer-Thurow, G. (1982). The industrialization of invention: a case study from the german chemical industry. *Isis*, *73*(3), 363–381.

Miller, S. P. (2010). Patent 'trolls': Rent-seeking parasites or innovation-facilitating middlemen? *Available at SSRN 1885538*.

Mowery, D. C., Sampat, B. N., & Ziedonis, A. A. (2002). Learning to patent: Institutional experience, learning, and the characteristics of us university patents after the bayh-dole act, 1981-1992. *Management Science*, *48*(1), 73–89.

Mowery, D. C., & Ziedonis, A. A. (2002). Academic patent quality and quantity before and after the bayh–dole act in the united states. *Research Policy*, *31*(3), 399–418.

Powers, J. B., & Campbell, E. G. (2009). University technology transfer: in tough economic times.

Change: The Magazine of Higher Learning, *41*(6), 43–47.

Rai, A. K., & Eisenberg, R. S. (2003). Bayh-dole reform and the progress of biomedicine. *Law & Contemp. Probs.*, *66*, 289.

Raj, A. (2004). The increasingly proprietary nature of publicly funded biomedical research. *Buying in or selling out*, 117–126.

Risch, M. (2015). A generation of patent litigation. *San Diego L. Rev.*, *52*, 67.

Rooksby, J. H. (2011). University initiation of patent infringement litigation. *John Marshall Review of Intellectual Property Law*, *10*(4), 623.

Rooksby, J. H. (2012). Innovation and litigation: tensions between universities and patents and how to fix them. *Yale JL & Tech.*, *15*, 312.

Rooksby, J. H. (2013). When tigers bare teeth: A qualitative study on university patent enforcement. *Akron L. Rev.*, *46*, 169.

Shane, S., & Somaya, D. (2007). The effects of patent litigation on university licensing efforts. *Journal of Economic Behavior & Organization*, *63*(4), 739–755.

Sherer, T., & Vertinsky, L. (2020). The innovation arms race on academic campuses. In *Research handbook on intellectual property and technology transfer.* Edward Elgar Publishing.

Shrestha, S. K. (2010). Trolls or market-makers-an empirical analysis of nonpracticing entities. *Colum. L. Rev.*, *110*, 114.

Squicciarini, M., Dernis, H., & Criscuolo, C. (2013). Measuring patent quality.

Sterzi, V., Pezzoni, M., & Lissoni, F. (2019). Patent management by universities: evidence from italian academic inventions. *Industrial and Corporate Change*, *28*(2), 309–330.

Thursby, J. G., Jensen, R., & Thursby, M. C. (2001). Objectives, characteristics and outcomes of university licensing: A survey of major us universities. *The journal of Technology transfer*, *26*(1-2), 59–72.

Tong, X., & Frame, J. D. (1994). Measuring national technological performance with patent claims data. *Research Policy*, *23*(2), 133–141.

US Federal Trade Commission, C. (2016). Patent assertion entity activity an ftc study.

Valdivia, W. D. (2013). University start-ups: Critical for improving technology transfer. *Center for Technology Innovation at Brookings. Washington, DC: Brookings Institution*.

Washburn, J. (2008). *University, inc.: The corporate corruption of higher education*. Basic Book

Measuring the impact of clinical data in terms of data citations by scientific publications

Yongmei Bai[1, 2] and Jian Du[2]

[1] *baiym2018@126.com*
School of Public Health, Peking University, No. 38 Xueyuan Road, Beijing (China);
National Institute of Health Data Science, Peking University, No. 38 Xueyuan Road, Beijing (China)

[2] *dujian@bjmu.edu.cn*
National Institute of Health Data Science, Peking University, No. 38 Xueyuan Road, Beijing (China)

Abstract

To explore the characteristics of highly cited data records, and exploring the relationship between data sharing/reuse of data records and the socio-economic burden of diseases / funding, the basic information of clinical data records were sorted. We extracted the subject words and exploring the diseases involved in the data records, so as to achieve the matching between the data records and diseases. We found that the data storage type of the DCI platform is data sets mainly, but the most cited data type is repositories. And the number of highly cited papers in data records is much lower than that of highly cited papers in the same field. The highly cited data records can reflect diseases of high concern to a certain extent. The disease related data records' sharing/reuse is positively correlated with its DALYs / foundation counts. Results shows that the data storage format should be further unified, the citation format should be further standardized. The data records on Chronic diseases and lifelong acquired diseases, such as Lower Respiratory Infections, Diabetes Mellitus and HIV/AIDS should be paid more attention.

Background

With the coming of the era of big data, more and more data records are measured and stored. The generation of massive data has brought unprecedented "data explosion". Almost every research field is eager for open access and use of data. Data citation is very necessary in order to protect the rights and interests of data providers and ensure the scientific and traceability of data. The practice of formal data citation includes data references and bibliographic references in the reference section of a publication. But the study showed that "the informal data citation in the main text of articles is far more common than formal data citations in the references of articles" (H. Park, You, & Wolfram, 2018). Management measures for data sharing and reuse are gradually implemented with more and more platforms for open data.

The current situation of data reference needs to study the data reference in specific data records. Informal reference data is difficult to track, but it is still an important way for researchers to share and reuse data, especially in the fields of life science and biomedicine field (H. Park & Wolfram, 2017). At the beginning of the 21st century, the genome science committee of Japan's Council for Science and Technology call for that "To assure scientific progress, all data and results from human genome research should be made public. (Triendl, 2000)" In August, 2013, there were some posters began appearing in doctor's practices in England to establish a unified management database of patients' electronic health records and use them for scientific research. The practice of sharing health records data is conducive to establish the National Health Service(NHS) system in England (Callaway, 2013). The Swiss National Science Foundation (SNSF) strongly supports the sharing of data and open access to data. They hope to achieve the findability, accessibility, interoperability and reusability of data through depositing the scientific data in any recognized digital archive (Egger & Kalt, 2017). In recent years, the application of open data records in the medical science field has helped the containment of infectious diseases, the research of cell and molecular biology, and the application of clinical drugs to solve medical problems. Data sharing is very important for all medical science research, especially for the outbreak of new and unexplored diseases. The publication of scientific paper cannot disclose relevant information at the first time, while the opening of original data can not only improve the value of shared data, but also help save valuable public health resources. Data sharing in medical science accelerates the process of disease research. For example, WHO shared data about Zika virus, and COVID-19 data set shared by Shanghai Public Health Clinical Center in January 2020 (Kieny, Moorthy, & Bagozzi, 2016; Zastrow, 2020).

The spirit of openness is becoming more and more popular in the scientific community, and many researchers and institutions call for making research data, software code and experimental methods publicly available and transparent (Gewin, 2016). Data reference is becoming more and more important in data sharing. However, the existing data are scattered in various industries and platforms, and have not been effectively integrated. There is a lack of unified management and reference format. It is very difficult to realize data sharing and unified use format. However, some data are related to data sets, data repositories or published research (H. Park & Wolfram, 2017). In order to explore the current situation of data sharing and reusing, we took the output and citations of clinical related data records in the Data Citation Index (DCI) platform as an observation. Focusing on clinical related data, we try to figure out 1) the distribution of data records types and related types of diseases in DCI, 2) the most highly cited data records, and 3) the relationship between data sharing/reuse and diseases-disability adjusted life years (DALYs) (WHO, 2018) for diseases, to compare global health needs and data sharing/reuse efforts. 4) the relationship between data sharing/reuse and diseases funding count, to demonstrate if there is an association between funding, and data sharing/reuse.

Methods

Data sources

Open access data records came from DCI database: DCI is a database intend to facilitate the discovery of data, link data to paper, and encourage citation of data. DCI on the Web of Science was selected because DCI provides a single access to over 500 data repositories worldwide and to over two million data studies and datasets across multiple disciplines and monitors quality research data through a peer review process (H. Park & Wolfram, 2017). A study showed that "Thomson Reuters developed this database in response to a stated desire among members of the research community for increased attribution of non-traditional scholarly output such as data records." DCI was launched in October of 2012 on the Web of science research platform. Cross-disciplinary search capabilities in the Index enable new possibilities for data discovery and synthesis. Evaluation of data records can be realized through DCI, and data records' relevance to scientific and scholarly research can be explored (Force & Robinson, 2014). While the reference of informal data records is not recorded in it, the platform can be used to explore the formal citations by scientific publications of data records (H. Park et al., 2018). As this study focuses on exploring the usage of clinical related data records, "Life Sciences & Biomedicine", one of the five research domains of WOS, is selected for retrieval in DCI platform. Among them, 57 subject categories have been obtained in this domain on August 7th, 2020, and 39 discipline categories related to Life Sciences & Biomedicine were obtained after 18 discipline categories without data were removed. In total, 13153 clinical data records were included in our study after removing the records related to molecular biology/pharmacy/animal. The workflow is shown in Figure 1.

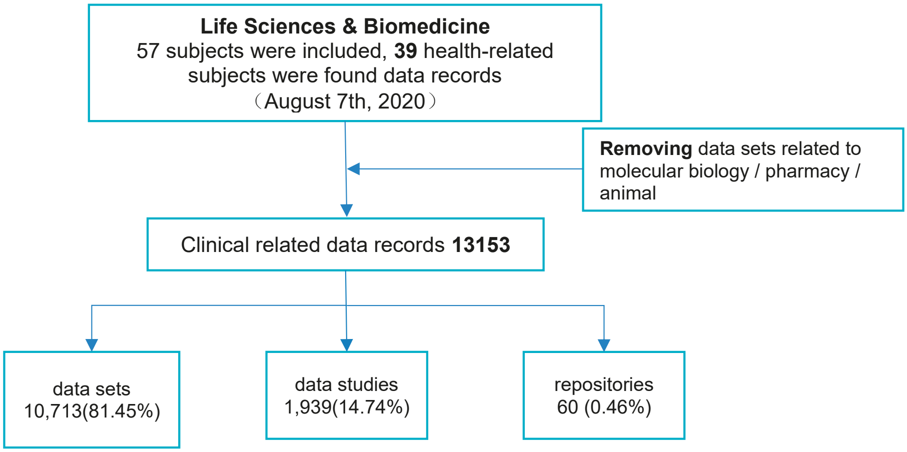

Life Sciences & Biomedicine
57 subjects were included, 39 health-related subjects were found data records
（August 7th, 2020）
Removing data sets related to molecular biology / pharmacy / animal
Clinical related data records 13153
data sets
10,713(81.45%)
data studies
1,939(14.74%)
repositories
60 (0.46%)

Figure 1. Data Records Filtering Flowchart

Data processing:

The Medical Text Indexer (MTI) produced by the National Library of Medicine (NLM) was adopted to extract standardized medical subject headings (MeSH) terms from the title and abstract of data records in DCI. We focused on the diseases related MeSH terms.

A correspondence table (Yegros-Yegros, van de Klippe, Abad-Garcia, & Rafols, 2020a) was used to map World Health Organization (WHO) International Classification of Diseases (ICD-10) with MeSH terms. The WHO Global Burden of Disease (GBD) survey provides useful information, including the corresponding codes of the ICD-10 and the global DALYs. The diseases distribution of DCI data records as well as their burdens was then explored.

In order to evaluate the relationship between the data sharing, data reuse and health demand at the disease level, the quantity of data records related to diseases (indicating the level of data sharing), the citations by scientific publications (indicating the level of data reuse, that is, using them to generate new research) and the social burden of diseases were analyzed respectively. DALYs, one of the Global Burden of Disease, GBD) indicators provided by WHO, is adopted as the index for disease burden assessment.

Concerning the funding for diseases, we noticed a data source in Global Observatory on Health R&D, which collected data on diseases/conditions in the number of grants for biomedical research by funder, type of grant, duration and recipients (WHO, 2019). This report provided statistics related to funding from 10 organizations. We collected data on the classification of disease/condition in the report as funding counts. SPSS 20.0 was used for statistical analysis of all the data. After single sample K-S test, the data of publications, citations and DALYs/funds of various diseases were in accordance with normal distribution, and Pearson Correlation Analysis was conducted to analyze the correlation among variables. Two-sided probability is taken as inspection level. P < 0.05 is statistically significant.

Results

Overall distribution of the data records

A total of 13153 clinical related data records were retrieved, which first appeared in 1900. Data records are mainly concentrated in the recent five years. According to the classifications in DCI, the data records are divided into three types: 10713 (81.45%) data sets, 1939 (14.74%) data studies and 60 (0.46%) repositories. Another 441 (3.35%) records are software. Among all the data records included in our study, 982 data records were cited, that is, new studies were produced through the reuse of these data records (Figure 2). But the most cited data record type is repositories.

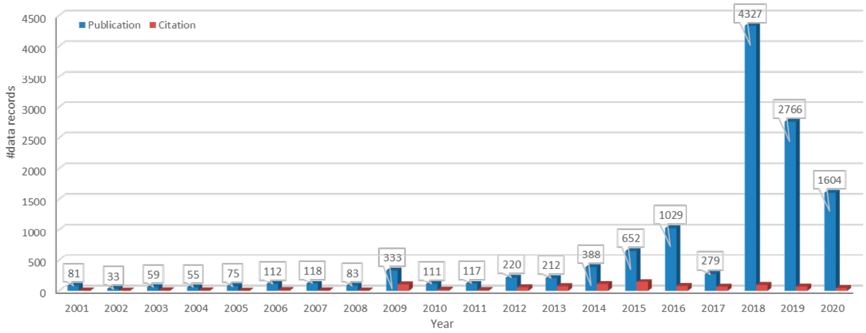


Figure 2. Number of diseases related data records in DCI

We could find the distribution of citations in clinical related data records according to table 1. Although the repositories related to clinical has the lowest proportion of data records, and their cited data records account has the lowest proportion. However, its citations by publications have the most frequency. To a certain extent, it shows that the sharing level of data sets and data studies is higher in clinical, and their scope of influence is wider. Meanwhile, the reuse level of data repositories is higher, which has a greater impact on new research.

Table 1. Distribution of citations in clinical related data records

Data records type	*Number of records (%)*	*Number of cited records*	*Cited times by scientific publications (%)*
data sets	10,713 (81.45%)	425	610 (5.37%)
data studies	1,939 (14.74%)	469	3,973 (34.96%)
repositories	60 (0.46%)	51	6,744 (59.35%)

Among them, 37 of the cited records are software, and the specific situation of this part is not statistically analyzed.

Top 20 related diseases by the number of data records and their citations

There are some differences for the top 20 related diseases by the number of data records and their citations respectively. Such diseases as chronic obstructive pulmonary disease, lower respiratory infections and Infectious bowel disease ranked in the top 10 output of data records, yet out of the top20 reuse of data records. To some extent, the data records related to the above three diseases are broadly shared, but their reuse is relatively poor. The number of shared data records of bipolar disorder and inflammatory bowel disease, which list within the top 10 reuse, but is not in top 20 number of sharing. In addition, the rate of cited data records for such diseases as alcohol use disorders (14, 51.85%), acute hepatitis C (10, 30.30%), brain and nervous system cancers (8, 21.05%) and breast cancer (6, 25.00%), are over 20%, whereas they are not in the top 20 output diseases.

Table 2. Top 20 related diseases by the number of data records and their citations

Rank	*All Records*		*High Cited Records*	
	Disease	*N*	*Disease*	*N (%)*
1	Diabetes mellitus	429	Schizophrenia	32 (18.82)
2	Trachea, bronchus, lung cancers	368	Trachea, bronchus, lung cancers	24 (6.52)
3	Asthma	342	Skin diseases	22 (7.91)
4	Chronic obstructive pulmonary disease	295	Melanoma and other skin cancers	21 (14.89)
5	Skin diseases	278	HIV/AIDS	20 (11.36)
6	Lower respiratory infections	236	Diabetes mellitus	20 (4.66)
7	Inflammatory bowel disease	210	Rheumatoid arthritis	19 (9.55)
8	Rheumatoid arthritis	199	Bipolar disorder	17 (19.77)
9	HIV/AIDS	176	Inflammatory bowel disease	16 (7.62)
10	Schizophrenia	170	Alzheimer's disease and other dementias	15 (14.29)
11	Epilepsy	159	Alcohol use disorders	14 (51.85)
12	Hypertensive heart disease	153	Epilepsy	13 (8.18)
13	Major depressive disorder	147	Acute hepatitis C	10 (30.30)
14	Melanoma and other skin cancers	141	Migraine	10 (10.87)
15	Colon and rectum cancers	141	Brain and nervous system cancers	8 (21.05)
16	Multiple sclerosis	136	Attention deficit/hyperactivity syndrome	8 (13.79)
17	Kidney diseases	120	Asthma	8 (2.34)
18	Alzheimer's disease and other dementias	105	Leukaemia	7 (7.22)
19	Acute hepatitis B	99	Gynecological diseases	7 (13.73)
20	Leukaemia	97	Breast cancer	6 (25.00)

The top 10 highly cited data records

Table3. The top 10 highly cited data records by scientific publications

Rank	*Title*	*Document Type*	*Group Author(s)*	*URL*	*Times Cited*
1	Longitudinal Aging Study Amsterdam	Repository	LASA Study Team	http://www.lasa-vu.nl	765
2	Swiss HIV Cohort Study	Repository	Swiss HIV Cohort Study Team	http://www.shcs.ch/	701
3	Avon Longitudinal Study of Parents and Children (ALSPAC)	Repository	ALSPAC Study Team	http://www.bristol.ac.uk/alspac/	574
4	Fragile Families and Child Wellbeing Study	Repository	Bendheim-Thoman Center for Research on Child Wellbeing; Columbia Population Research Center	http://www.fragilefamilies.princeton.edu	480
5	Nationwide Inpatient Sample (NIS)	Data study	Agency for Healthcare Research & Quality (AHRQ) Healthcare Cost and Utilization Project (HCUP)	http://hcup-us.ahrq.gov/nisoverview.jsp	383
6	Study of Health in Pomerania	Repository	SHIP Organization Center Study of Health in Pomerania	http://www.medizin.uni-greifswald.de/cm/fv/english/ship_en.html	322
7	Norwegian Women and Cancer Study	Repository	NOWAC Study Team Norwegian Women and Cancer Study	http://site.uit.no/nowac/	305
8	Health Improvement Network	Repository	Cegedim Strategic Data Medical Research UK Health Improvement Network	http://www.thin-uk.com/	286
9	Kids' Inpatient Database (KID)	Data study	Agency for Healthcare Research & Quality (AHRQ) Healthcare Cost and Utilization Project (HCUP)	http://hcup-us.ahrq.gov/kidoverview.jsp	271
10	Nord-Trondelag Health Study	Repository	HUNT Research Centre Nord-Trondelag Health Study	http://www.ntnu.edu/hunt	269

Most of the top 10 cited data records are longitudinal cohort studies, which are often carried out for specific groups or regions. Among the top 10 highly cited data records, 8 data records are repositories, and the other 2 are data studies. The most cited data record among them is Longitudinal Aging Study Amsterdam, and it has been officially cited by 765 scientific publications at present (Table 3). Van Noorden, R et al. published a study about the top 100 papers in the world in 2014, which showed that "the number of cited papers ranked first in bioinformatics was 305,148", and the number of cited papers ranked tenth was 40,289" (Van, Maher, & Nuzzo, 2014). It is observed that the cited frequency of data sets is generally lower than that of scientific papers.

The relationship between data sharing/reuse and DALYs

According to our statistics, the published data records (data sharing) were related to 114 diseases and directly linked to 6032 data records. The citations of data records involved 51 diseases and included 369 data records. Pearson correlation analysis was made between the values of Publications and Citations and DALYs in SPSS20.0. The results showed that there was a significant correlation between the two groups of data, and the correlation coefficients were 0.341 P<0.001) and 0.985 (0.003), respectively (Table 4). To some extent, there is a positive correlation between the sharing/reuse of data sets and the disease related DALYs.

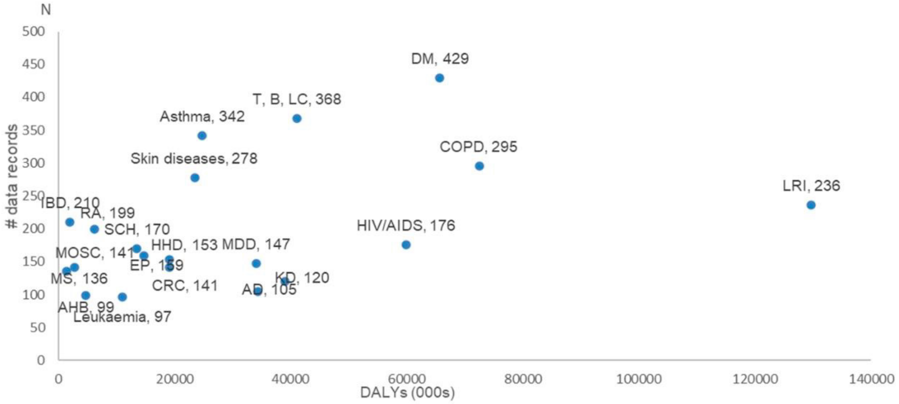
Table4. The relationship between data sharing/reuse counts and DALYs/Funding counts

	r	*P*
# Data sharing vs. DALYs	0.341	<0.001
# Data reuse vs. DALYs	0.985	0.003
# Data sharing vs. Funding counts	0.428	<0.001
# Data reuse vs. Funding counts	0.375	0.020

The published and cited data records involved 77 and 38 diseases respectively, which also have the data of funding count in the World report. We made Pearson correlation analysis between the number of publications, citations and diseases' funding counts. The results showed that there was also a significant correlation between the two groups of data, and the correlation coefficients were 0.428 (<0.001) and 0.375 (0.020), as is shown in Table 4. We tend to conclude that the level of data sharing/reuse for various diseases might be promoted by their funding counts.

The relationship between top 20 shared/reused data records and DALYs

We have found that the total number of disease-related data records and that has been cited by publications were positively correlated with their DALYs overall. In the top 20 number of disease related data sets, lower respiratory infections have the highest DALYs (129689.84), and then is followed by chronic obstructive pulmonary disease (72512.38), diabetes mellitus (65666.15) and HIV/AIDS (59951.10).

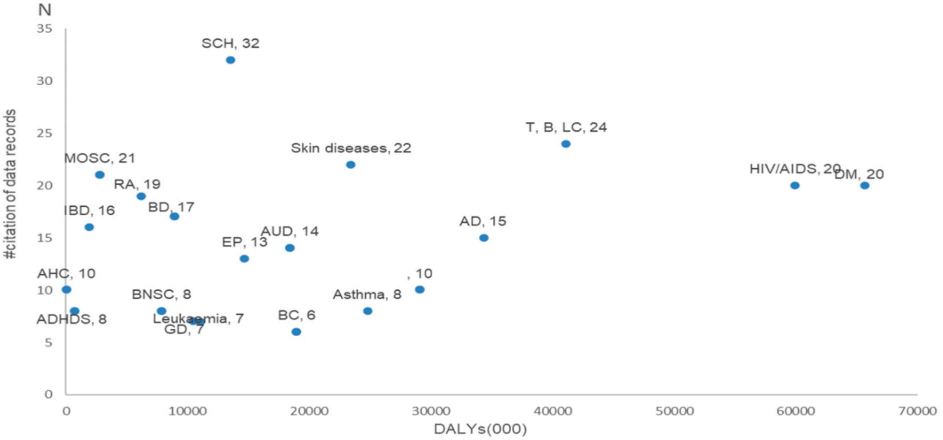


DM: Diabetes mellitus; T, B, LC: Trachea, bronchus, lung cancers; COPD: Chronic obstructive pulmonary disease; LRI: Lower respiratory infections; IBD: Inflammatory bowel disease; RA: Rheumatoid arthritis; SCH: Schizophrenia; EP: Epilepsy; HHD: Hypertensive heart disease; MDD: Major depressive disorder; MOSC: Melanoma and other skin cancers; CRC: Colon and rectum cancers; MS: Multiple sclerosis; KD: Kidney diseases; AD: Alzheimer's disease and other dementias; AHB: Acute hepatitis B

Figure 3. The relationship between the top 20 number of disease related data records and its DALYs

Diabetes mellitus has the highest number of data records among them. The number of lower respiratory infections related data records is relatively fewer compared with their DALYs (Figure 3).

In the top 20 number disease related data sets cited by publications, diabetes mellitus (65666.15) has the highest DALYs, and then is HIV/AIDS (59951.10) and trachea, bronchus, lung cancers (41121.34). Schizophrenia has the greatest number of cited data records, while its DALYs is not relatively high (13541.00) (Figure 4).

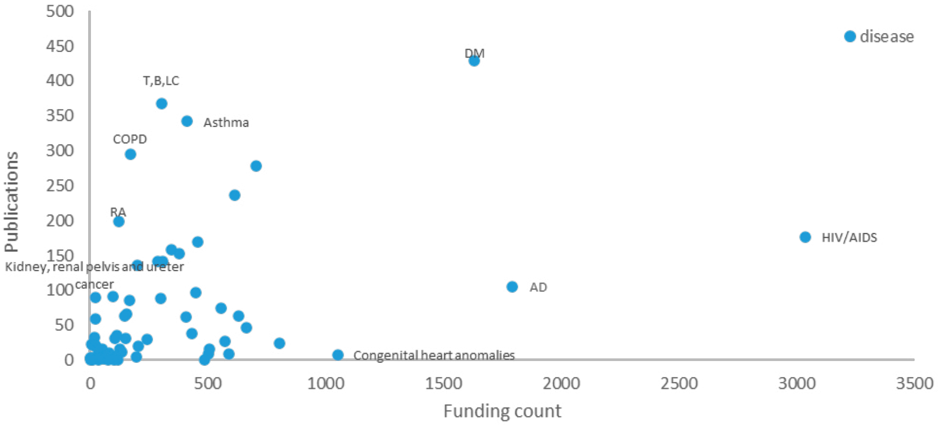


SCH: Schizophrenia; T, B, LC: Trachea, bronchus, lung cancers; MOSC: Melanoma and other skin cancers; DM: Diabetes mellitus; RA: Rheumatoid arthritis; BD: Bipolar disorder; IBD: Inflammatory bowel disease; AD: Alzheimer's disease and other dementias; AUD: Alcohol use disorders; EP: Epilepsy; AHC: Acute hepatitis C; BNSC: Brain and nervous system cancers; ADHDS: Attention deficit/hyperactivity syndrome; GD: Gynecological diseases; BC: Breast cancer

Figure 4. The relationship between the top 20 number of cited disease related data records and its DALYs

The relationship between the number of shared/reused data records and funding counts

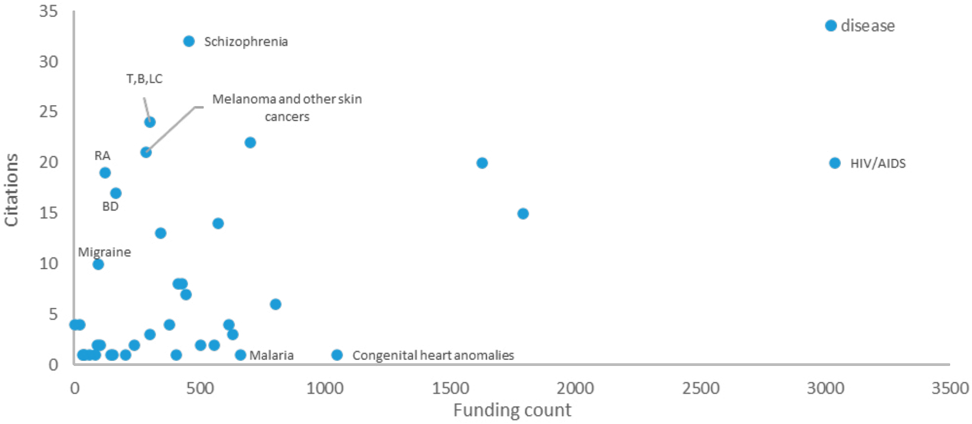
We have found that the total number of disease-related data records and that has been cited were positively correlated with their funding counts overall (Figure 5).



DM: Diabetes mellitus; T, B, LC: Trachea, bronchus, lung cancers; COPD: Chronic obstructive pulmonary disease; AD: Alzheimer's disease and other dementias

Figure 5. The relationship between the output of disease-related data records and its funding counts

For example, the diabetes mellitus ranked first by the number of shared data records and raked third by funding counts. But we also could see some diseases have fewer funding counts, but have a high level of data sharing, such as COPD, RA, asthma, kidney renal pelvis and ureter cancer and T, B, LC. The sharing of these data records may come from the interest of researchers, or it may be demand driven. But there are some diseases has a larger ratio of funding counts to data records sharing, such as HIV/AIDS, AD and congenital heart anomalies. The total number of disease-related data records and that has been reused were also positively correlated with their funding counts overall (Figure 6). But some diseases have a larger ratio of reused number to its funding counts, such as schizophrenia, melanoma and other skin cancers, RA, BD, migraine and T, B, LC. Some diseases have a larger ratio of funding counts to data records sharing, such as HIV/AIDS, AD and congenital heart anomalies.



T, B, LC: Trachea, bronchus, lung cancers; RA: Rheumatoid arthritis; BD: Bipolar disorder

Figure 6. The relationship between the citations of disease-related data records and its funding counts

Conclusions

Our study revealed that data sets are the main type of disease-related data records on DCI platform, followed by data studies and repositories. The overwhelming majority of data records are published in the recent five years, and show a trend of rapid growth year by year. However, the citations of data records are less than that of scientific publications, which may be related to the diversity of data records formats and the inconsistency of reference standards or norms, indicating that the standardized use of data (such as formal citation) needs to be further strengthened.

The mantra of the nascent open-data movement is that “scientists should share online all data underlying their findings”. It sounds simple, but it can be tough to achieve in practice. Many authoritative academic institutions in many countries have successively introduced and implemented various measures and policies to promote data openness and sharing. In 2014, the Gates Foundation announced that it would "require its funded researchers to disclose data", while some publications only encourage scientists to open their data (Van Noorden, 2014). The right to force researchers to abide by the rules of data use lies in the hands of editors and reviewers of journals, and magazines can formulate data citation policies to ensure that the used data is used in a standardized way (Roche, 2016). For example, *Science* and *Nature* have announced the policy of citing research data records in their journals (Hyoungjoo Park &

Wolfram, 2019). It shows that formulating data citation rules through scientific research funds and publications may have the effect of standardizing data formal citation to a certain extent.

According to the analytic data of data records sharing and reuse, the number of data records sharing for the following diseases can be appropriately increased in future research: Bipolar disorder, Inflammatory bowel disease, Alcohol use disorders, Acute hepatitis C, Brain and nervous system cancers and Breast cancer.

Highly cited data records may indicate diseases of high concern to a certain extent. Through correlation analysis of the number of data records corresponding to diseases, it can be found that the total number of disease-related data records and their citations by publications is positively correlated with DALYs as a whole. And more attention should be paid to Lower Respiratory Infections, Diabetes Mellitus and HIV/AIDS, which have a high disease burden.

Previous studies have shown that the burden of disease is not always strongly correlated with the amount of funds and scientific publications related to disease, for example, some underfunded or overfunded diseases can be found. And underactive or overactive disease areas can be also found (Yegros-Yegros, van de Klippe, Abad-Garcia, & Rafols, 2020b; Zhang, Zhao, Liu, Sivertsen, & Huang, 2020). We believe that the reason is that scientific research is driven by both interest and demand. Compared with the amount of funds and scientific publications, the sharing and reuse of data records should be more demand-driven, which mainly comes from disease burden or unmet clinical needs. We will verify it in future research for this hypothesis.

We analyzed the relationship between the data records sharing/reuse of diseases and the its funding counts. They have a significant positive correlation on the whole. However, some diseases deviate from this trend. Such diseases have many funding support but only can search a few data records in DCI, like HIV/AIDS, congenital heart anomalies, malaria and so on. This may because these diseases have a better prevention or control measure. The main purpose of funding is to promote some prevention or control activities, such as AIDS publicity and self-inspection materials distribution. Not scientific research. But some diseases have many data records but only a little funding support, such as diabetes mellitus, COPD, AD, asthma, and migraine. Most of this kind of diseases are chronic diseases. They do not have a good standard of care at present, and this may be why they have more data records than other diseases. And they may need more funding to have in-depth research. We need to explore the research field of highly cited data records in future research, so as to explore how these data records are used many times in subsequent research, such as whether important medical evidence is formed or even written into clinical practice guidelines, so as to evaluate the value of medical data.

Considering the limitations of DALYs, more indicators of disease burden should be included in future research to explore the research focus and development trend of health data records.

At present, the data on funding counts considered in our study only include 10 funding organizations. We will consider including the funding support data for more countries and international organizations to our study in the future research.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (71603280, 72074006), and the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (2017QNRC001).

References

Callaway, E. (2013). UK push to open up patients' data. *Nature, 502*(7471), 283-283. doi:10.1038/502283a

Egger, M., & Kalt, A. (2017). Open data: support from Swiss funder. *Nature, 547*(7664), 403-403. doi:10.1038/547403b

Force, M. M., & Robinson, N. J. (2014). Encouraging data citation and discovery with the Data Citation Index. *Journal of Computer-Aided Molecular Design, 28*(10), 1043-1048. doi:10.1007/s10822-014-9768-5

Gewin, V. (2016). DATA SHARING An open mind on open data. *Nature, 529*(7584), 117-119. doi:10.1038/nj7584-117a

Kieny, M. P., Moorthy, V., & Bagozzi, D. (2016). Use open data to curb Zika virus. *Nature, 533*(7604), 469-469. doi:10.1038/533469b

Park, H., & Wolfram, D. (2017). An examination of research data sharing and re-use: implications for data citation practice. *Scientometrics, 111*(1), 443-461. doi:10.1007/s11192-017-2240-2

Park, H., & Wolfram, D. (2019). Research software citation in the Data Citation Index: Current practices and implications for research software sharing and reuse. *Journal of Informetrics, 13*(2), 574-582. doi:10.1016/j.joi.2019.03.005

Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology, 69*(11), 1346-1354. doi:10.1002/asi.24049

Roche, D. (2016). Open data: policies need policing. *Nature, 538*(7623), 41-41. doi:10.1038/538041c

Triendl, R. (2000). Japan calls for open access to human genome data. *Nature, 405*(6784), 265-265. doi:10.1038/35012776

Van Noorden, R. (2014). Confusion over open-data rules. *Nature, 515*(7528), 478-478. doi:10.1038/515478a

Van, N. R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature, 514*(7524), 550.

WHO. (2018). *Global Health Estimates 2016: Disease burden by Cause, Age, Sex, by Country and by Region, 2000-2016.* Retrieved from Geneva:

WHO. (2019). Global Observatory on Health R&D. Retrieved from https://www.who.int/research-observatory/monitoring/inputs/world_report/en/

Yegros-Yegros, A., van de Klippe, W., Abad-Garcia, M. F., & Rafols, I. (2020a). Exploring why global health needs are unmet by research efforts: the potential influences of geography, industry and publication incentives. *Health Research Policy and Systems, 18*(1), 47. doi:10.1186/s12961-020-00560-6

Yegros-Yegros, A., van de Klippe, W., Abad-Garcia, M. F., & Rafols, I. (2020b). Exploring why global health needs are unmet by research efforts: the potential influences of geography, industry and publication incentives. *Health Research Policy and Systems, 18*(1). doi:10.1186/s12961-020-00560-6

Zastrow, M. (2020). OPEN SCIENCE TAKES ON COVID-19. *Nature, 581*(7806), 109-110. doi:10.1038/d41586-020-01246-3

Zhang, L., Zhao, W., Liu, J., Sivertsen, G., & Huang, Y. (2020). Do national funding organizations properly address the diseases with the highest burden?: Observations from China and the UK. *Scientometrics, 125*(2), 1733-1761. doi:10.1007/s11192-020-03572-9

An empirical analysis of existence of Power Laws in social media mentions to scholarly articles

Sumit Kumar Banshal[1], Aparna Basu[2], Vivek Kumar Singh[3], Hiran H Lathabai[4], Solanki Gupta[5] and Pranab K. Muhuri[6]

[1] *sumitbanshal06@gmail.com*, [6]*pranabmuhuri@gmail.com*
Department of Computer Science, South Asian University, New Delhi (India)

[2] *aparnabasu.dr@gmail.com*
Formerly with CSIR-NISTADS, New Delhi (India)

[3]*vivek@bhu.ac.in*, [4] *hiranhl007@gmail.com*, [5]*solankigupta2@gmail.com*,
Department of Computer Science, Banaras Hindu University, Varanasi (India)

Abstract

Power laws are a characteristic distribution found in both natural as well as in man-made systems. Previous studies have shown that citations to scientific articles follow a power law, i.e., the number of papers having a certain level of citation x are proportional to x raised to some negative power. However, the distributional character of altmetrics (such as reads, likes, mentions, etc.) has not been studied in much detail, particularly with respect to existence of power law behaviours. This article, therefore, attempts to do an empirical analysis of altmetric mention data of a large set of scholarly articles to see if they exhibit power law. The individual and the composite data series of ‘mentions’ on the various platforms are fit to a power law distribution, and the parameters and goodness of fit determined using least squares regression. We also explore fit to other distributions like the log-normal and Hooked Power Law. Results obtained confirm the existence of power law behaviour in social media mentions to scholarly articles and we conclude that altmetric distributions also follow power law with a fairly good fit over a wide range of values.

Introduction

A power law is simply expressed by the fact that if the probability distribution of measuring a quantity x varies as the inverse power of the same variable raised to a value k, $p(x) \sim x^{-\alpha}$, then x is said to follow a power law with exponent $-\alpha$ (Newman, 2007). Power laws were first noticed a century ago by Pareto in the context of income distribution (Pareto, 1896). They abound in nature as well as in the artificially created technology systems, most recently the Internet (Adamic, 2000; Faloutsos et al., 1999). Basically, any heavily tailed distribution could follow or model into the power law. This specific feature was observed in different disciplinary areas at different periods of time, over the last century by different people, and early power laws go by different names like the Bradford law (Bradford, 1934), Zipf's law (Zipf, 1949), Pareto’s law (Chen & Leimkuhler, 1986; Pareto, 1896). All these types of distribution models are said to imitate the Lotka’s Law (Lotka, 1926). Power law distributions involve various signature characteristics such as - if the full range is represented on the axes, the shape of the curve would be a perfect L. Also, if same distribution is plotted on log-log scale, then the curve is always linear in nature (Adamic, 2000). Also, the power law distribution is the only distribution that holds the scale–free property, i.e., the shape of distribution remains unchanged whatever be the scale at which it is observed (Newman, 2007). Egghe (2005) & Egghe (2009) explored power laws in information production processes (IPPs) and derived various properties associated with the power law in general IPPs and Lotkaian IPPs (where power law exponent $\alpha > 2$). Rousseau (1997) presented a webometric analysis perspective, wherein Lotka’s law was discussed with an indication of Power Law behaviours.

In the domain of scholarly articles, citations have been considered as a widely known parameter of impact assessment for a long time. Several previous studies have shown that Power laws are observed in the distribution of citations to scientific papers (Brzezinski, 2015; Redner, 1998, 2005; Thelwall & Sud, 2016). With the evolution of social media platforms and their increased use in scholarly publishing domain in the recent times, a new and quick event based impact measure has emerged, known as alternative metrics or 'altmetrics' (Priem et al., 2010; Priem & Hemminger, 2010). Many recent studies on altmetrics have pointed that they correlate with citations in different degrees (Banshal et al., 2021; Costas et al., 2015; Eysenbach, 2011; Haustein et al., 2014; Peoples et al., 2016; Shema et al., 2014; Thelwall, 2018; Thelwall & Nevill, 2018). Costas et al. (2016) applied the Characteristic scores and scales (CSS) method devised by Schubert et al. (1987) for selected altmetric platforms. That method classified articles into three classes and studied how those articles were distributed in three classes in different altmetric platforms. They also studied the implications of specific distribution in each platform for development of field-normalized indicators. Some studies, that analysed altmetric data of large sample of scholarly articles, have indicated that they are skewed in nature (Thelwall & Nevill, 2018; Thelwall & Wilson, 2016). However, the altmetric data distributions are not analysed in detail with respect to existence of power laws.

This article analysed the social media mention data drawn from four different platforms (Twitter, Facebook, News and Blog) for a large set of scholarly articles obtained from Web of Science database. Our objective was to test if altmetric indicators also follow power laws – as seen in the case of citations, the Internet, and numerous natural and man-made systems. The altmetric mention data for articles was visualized in log-log space. The individual and the composite data series of 'mentions' on the various platforms are fit to a power law distribution, and the parameters and goodness of fit determined using least squares regression. We also explore fit to other distributions like the log-normal and Hooked Power Law. The article attempts to answer following key research questions:

1. Do altmetrics indicators (mentions) follow Power Laws?
2. Will other distributions like hooked power law and discretised lognormal distribution be plausible for this data?

Related Work

Power law properties are ubiquitous in various natural and behavioural phenomena. Power laws are seen to appear in large, interconnected and self-organising systems. It has scale free nature and has applications in various domains like physics (e.g. sandpile avalanches), economics, linguistics, biology (e.g. species extinction), finance, information and computer science, geology, social science, astronomy etc. (Clauset et al., 2009; Newman, 2007). A lot of studies have been done regarding the emergence of power laws in diverse fields. Power Laws help in explaining various phenomena of Science (Mitzenmacher, 2004), Economics (Gabaix, 2009, 2016), Financial Time series (Bouchaud, 2001), and also network topologies (Faloutsos et al., 1999). The statistics for number of visitors to sites of the World Wide Web and distribution of visitors per site observed, and it approximated a power law (Adamic, 2000; Adamic & Huberman, 2001). There is a long list available as a proof of occurrence of power laws in various disciplines. Some of them mentioned in Newman (2007) are- word frequency count in texts, citations of scientific papers, web hits, copies of books sold, telephone calls, magnitude of earthquakes, diameter of moon craters, intensity of solar flares, intensity of wars, wealth of the richest people, frequencies of family names, population of cities and so on.

The first mention about the existence of power laws in the scholarly articles was made by Solla Price (Price, 1965) for the highly cited articles. Later on, he recommended 'cummulative advantage' mechanism which also can cause power law distribution (Price, 1976). Redner (1998) analyzed the citation distribution of scientific publications and found that the asymptotic tail of the citation distribution appears to be described by a power law, with exponent equal to 3. In another work, Redner analysed articles published over a century in the journal of physical review and found the tailed characteristics of citations (Redner, 2005). The exponent factor of highly cited papers and relatively lesser cited papers found to vary as studied by (Peterson et al., 2010). They analysed two types of mechanisms in citations namely 'direct' and 'indirect' mechanism. In this approach, three different set of scientific articles were examined. The existence of power laws in citations are relatively universally found in the context of research discipline (Radicchi et al., 2008). In a more comprehensive and elaborate empircal approach, Brzezinski (2015) detected power-law behaviour in the citation distribution. In this work, scientific papers published between 1998 and 2002 were drawn from Scopus and analysed. They found that the power-law hypothesis is not satisfied for around half of the Scopus fields of science.

A more robust and quantitative analysis of statistical distributions was performed by Thelwall (2016c) by considering data from 26 Scopus subject areas and seven years including 911,971 journal articles. The study considered three different models-the Hooked Power law model, the Truncated Power Law model, and the Discretised log normal model. He tested whether the discretised lognormal and hooked power law distributions were plausible for citation data. It was also tested if there were too many uncited articles, and zero inflated variants of the discretised lognormal and hooked power law distributions (Thelwall, 2016b). He examined best options for modelling and regression, and tried to detect which distribution best fit citation data by including and excluding uncited articles (Thelwall, 2016a). He concluded that the hooked power law and discretised lognormal distribution were the best options for complete citation data. Though, several studies (Costas et al., 2015a; Haustein et al., 2014; Eysenbach, 2011; Mohammadi et al., 2016; Thelwall, 2018; Thelwall & Nevill, 2018) have found positive correlations between citations and altmetrics, but there are just few studies on distribution patterns of altmetric data.

The only major studies on distribution analysis of altmetric data include analysis of online events- web usage counts (Wang et al., 2016), readership (Thelwall & Wilson, 2016) and download patterns (Duan & Xiong, 2017). Wang et al. (2016) investigated the Web of Science usage count to understand the distribution and the relation with citation counts. They analysed data for about 12,000 articles from five journals of information science domain. The 'usage count' for each article was analysed for fit with power law distribution and positive evidence of power law behaviours was observed for usage counts. The study by Thelwall & Wilson (2016) analysed about 332,000 scientific publications in 45 subfields of medical sciences domain drawn from Scopus database. The study observed the relationship between Mendeley reads and citations and also examined the distribution patterns of reads by exploring fitsto either lognormal or hooked power law. They concluded that, hooked power law is a better fit in case of citation data, but for Mendeley reads, it varied across 45 sub-fields. Duan & Xiong(2017) analysed the download patterns from the Chinese Library & Information Sciences using two-step cluster analysis, outlining the distribution for the same. The authors collected data from eleven chinese journals and grouped them into clusters based on the type of downloads. They concluded that the majority of sets followed power function in the case of absolute downloads. Costas et al. (2016) studied distribution patterns of citations and altmetrics (Mendeley reads, tweets, blogs) and found that twitter and blog mentions are extremely skewed distributions. They observed that it was largely caused by a large share of publications without any mentions on these platforms. There are, however, no previous studies on examining the distribution patterns of mentions in many popular social media platforms like Twitter, Facebook, Blogs etc., particularly with respect to Power Law behaviours.

Data

The article obtained the data for analysis from two sources: Web of Science and Altmetric.com. The scholarly article data was obtained from Web of Science (WoS) and the altmetric mention data for these articles was obtained from Altmetric.com.

The scholarly article data downloaded comprised for the whole Web of Science publication records for the year 2016. There was a total of 2,528,868 publication records for which the metadata was downloaded. The data was download in the month of Sep. 2019 and comprised of all document types. Out of the 2,528,868 records, only 1,785,149 publication records were found to have a DOI (Digital Object Identifier). We have, therefore, used the data for these 1,785,149 publication records for further analysis. This was done due to the fact that DOI was the link between Web of Science and Altmetric.com data. The altmetric data for these publication records was collected from the Altmetric.com aggregator through an automated DOI-based lookup process. The Altmetric.com aggregator accumulates social and online mentions of research output for more than 18 different social and online platforms (such as Twitter, Facebook, Blog, News, Mendeley etc.). Out of the 1,785,149 publication records, altmetric mention data for a total of 902,990 publication records was found and downloaded. **Table 1** provides the details of data. The data obtained from Altmetric.com had 46 fields, including DOI, title, Twitter mentions, Facebook mentions, News mentions, altmetric attention score, OA Status, Subjects (FoR), publication date, URI, etc. We have used the altmetric attention score and mention data for Twitter, Facebook, News and Blog for the analysis.

Table 1. Data collected from Web of Science and Altmetrics.com

Total WoS records (2016)	*WoS records with DOI*	*Publication records with Altmetric.com data*
2,528, 868	1,785,149	902,990

Methodology

For answering the first research question posed, the standard strategy for revealing a power law is to draw a size-frequency plot, i.e., to plot the number of papers N(k) with a given level of mentions k. In other words, we use the fact that a histogram of a quantity with a power law distribution appears as a straight line when plotted on a logarithmic scale (Newman, 2005). Therefore, the mentions were plotted using log-log plot and the parameters of fit calculated to ascertain the plausibility of a power-law. The plots were made for altmetric attention score and mention data for Twitter, Facebook, News and Blog platforms.

In order to answer the second research question, we have tried the following three models, which may all be characterised as empirical models: the truncated power law, the lognormal, and the hooked power law (Thelwall, 2016c, 2016a). Since the altmetrics data is discrete, while the lognormal is for continuous data we have used the Discretized lognormal distribution and tested them for goodness of fit against data taken from Altmetric.com.

Lognormal Distribution: It is a continuous probability distribution that fits data when the natural logarithm of the data follows a normal distribution. Since natural logarithm cannot be calculated for zero or negative numbers, hence only positive values are allowed for such distributions. The lognormal distribution is described as,

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\frac{\ln(x-\mu)^2}{2\sigma^2})} . \qquad (1)$$

It has two standard parameters, the mean (μ) and standard deviation (σ). Modifications are required in order to convert it into discrete form for discrete data. An offset 1 is added to the mentions for such calculation giving the ***discretized lognormal distribution***.

Hooked power law distribution: There are several other methods of conversion from continuous to discrete distribution (Thelwall, 2016c). The *hooked power law* is defined as,

$$f(x) = A * \frac{1}{(B+x)^{\alpha}} \qquad (2)$$

where an *ad hoc* parameter B is added to the variable x. Here the two parameters that affect the process are "A" and "B". Setting the offset parameter "B" to zero will lead to the pure form of the power law.

Truncated Power Law: A better fit to data can be obtained by simply truncating the data series at the tail end. Again the process is ad hoc. Since the tail of the distribution is sparsely populated, the data loss may not be high. However since the fluctuations are high in the tail region, by truncation we effectively reduce fluctuations and obtain a better model fit to data. Various multiplicative models are also used to generate such distributions. Only a small change from the generative process of one, leads to the other distribution. The power law gives a good visual fit but when logarithms are used the relationship between the raw variables may not be as accurate.

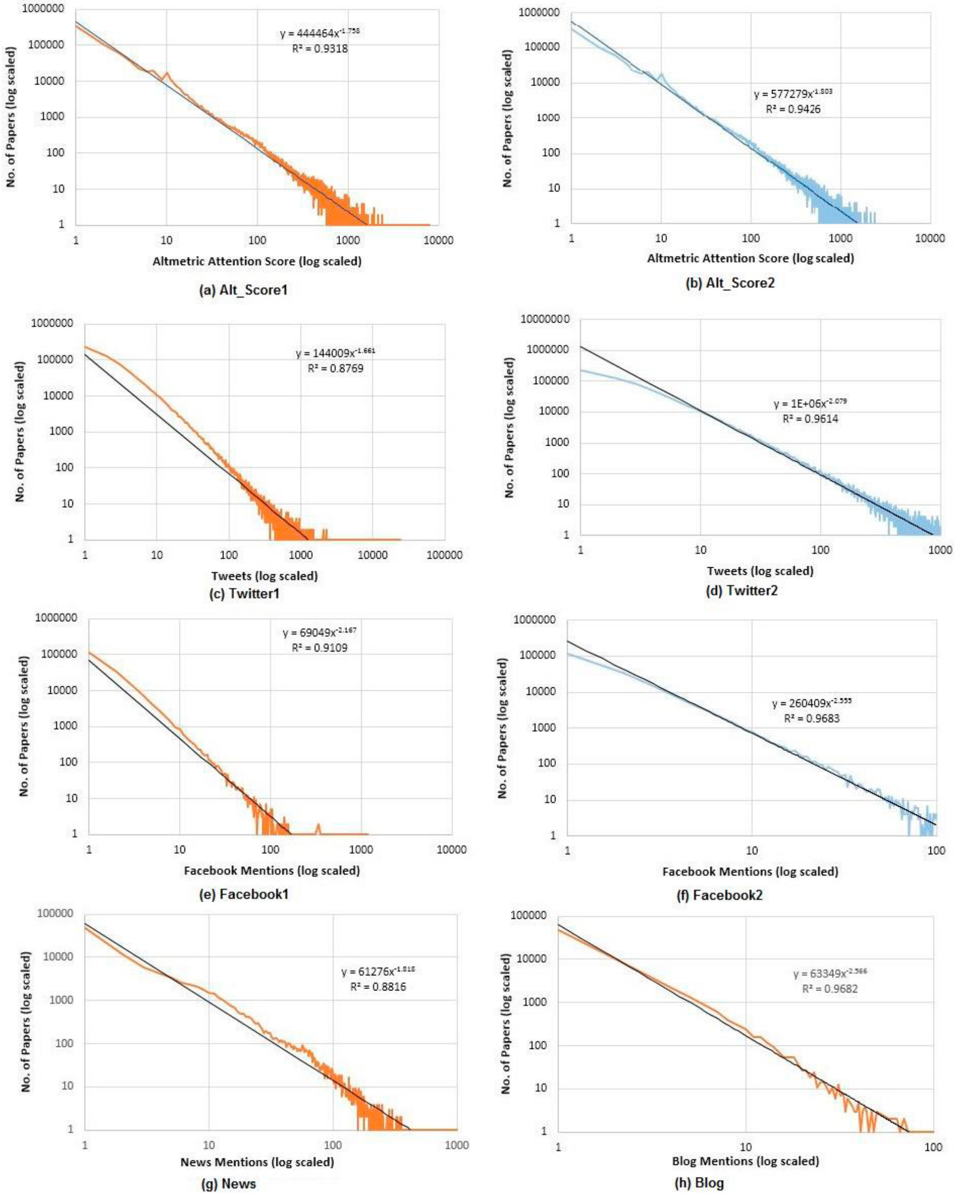
Results

In order to understand the distribution of the data, the four prominent platforms in Altmetric.com data along with the composite score have been explored, namely, Facebook, Twitter, Blogs, News and Alt-score (a combined index created out of all the other indicators). The mentions were plotted using log-log plot and the parameters of fit calculated to ascertain the plausibility of a power-law. We find that the altmetrics data when plotted show reasonably good linear fit (Figure 1). The figure shows plot for the full as well as truncated data. In each case, supplement '1' corresponds to full data and supplement '2' corresponds to truncated data. The power ranges between -1.661 to -2.614, and the variance squared explained by the model ranges between $0.8769 < R^2 < 0.9683$.

Truncated Power Law

It is easy to see from the plots (Figure 1) that the fit may look good, but the parameters are not too high ($R^2 <\sim 1$). $R^2 = 0.9682$ for Blogs, which gives the best fit to a power law. The least value of R^2 is 0.8769 for Twitter with poorest fit (Table 2).

Can the degree of fit of the Power Law be improved? By truncating the distribution in the tail, it is possible to obtain a better fit. It can be seen from the plots that the fit becomes poor at high values of accumulated mentions. In other words, there are only a few papers that merit a high rate of mention, and therefore there are statistical fluctuations at this end. By truncating the distribution at the high frequency end, we get an improvement in the degree of fit, without losing too many data points. Truncation was tried only with the Alt-score, Twitter and Facebook variables, as the data series for News and Blogs were short and therefore not suitable for truncation. The increase in the parameter of fit, R^2 was 1.2%, 9.6% and 6.3 %, respectively, for the Alt-score, Twitter and Facebook (Table 2). The truncated graphs are labelled with the subscript ‘2’ for each of the variables. (Figure 1, Table 2).



(a) Alt_Score1

(b) Alt_Score2

(c) Twitter1

(d) Twitter2

(e) Facebook1

(f) Facebook2

(g) News

(h) Blog

Figure 1. Power Law Fit to Altmetric Data: Altmetric data plotted on log-log scale with linear fit, parameters, and variance explained.

Table 2. Parameters of Fit to a power law and truncated power law for altmetric data

Data Source	*Data points*	*A*	*B*	*R^2*	*% change*
Alt_Score1	*757784*	*444464*	*-1.758*	*0.9318*	*-*
Alt_Score2	*757687*	*577279*	*-1.803*	*0.9426*	*1.2*
Twitter 1	*700985*	*144009*	*-1.661*	*0.8769*	*-*
Twitter2	*700714*	*1E+06*	*-2.079*	*0.9614*	*9.6*
Facebook1	*185243*	*69049*	*-2.167*	*0.9109*	*-*
Facebook2	*185167*	*260409*	*-2.555*	*0.9683*	*6.3*
News	*98499*	*61276*	*-1.818*	*0.8816*	*-*
Blog	*70387*	*63349*	*-2.614*	*0.9682*	*-*

Note: In each case, supplement '1' corresponds to full data and supplement '2' corresponds to truncated data.

Altmetric data fit to other distributions: Hooked Power Law and Discretized Lognormal

As we have seen truncation of the data series can improve the goodness-of fit of the Power law model to altmetrics data. However, it can be argued that we are essentially throwing away data in order to get a good fit. Neither can it be theoretically specified which range of data to discard. There could be other distributions, other than the Power law, which give an equally good fit, or even a better fit. These distributions need to be tested against the data. We try two alternative distributions discussed by Thelwall (2016c).

An alternative to assuming that the distribution of the altmetric variables is a power law, or close to a power law, would be to look for other distributions that give a close fit to the data, this is the second research question we have noted. For addressing the second research question, the plausibility of hooked and discretised lognormal power law in altmetric data is tested. The hooked power law concept is based upon the impression that citation can occur in two processes. One of them is that citations can be accrued randomly and in the other concept, one article is getting cited due to its' earlier citations. Merging these two phenomena then the probability that an article attracts *k* citations is proportional to $1/ (B + k)^{\alpha}$. When the offset parameter, B converges to zero than it shows the distribution of pure power law (Thelwall, 2016c).

Two bootstrapped samples were drawn from all four altmetric sources (Twitter, Facebook, Blog and News) to examine the same. As mentioned by Clauset et al. (2009) if data is compatible with power law distribution, then the data is also prone to be comparable with other empirical distributions like Discretised Lognormal, Hooked and Stretched Exponential distributions. Therefore, we checked the plausibility of Hooked and discretised lognormal distributions using standard statistical techniques. The statistical approach that is followed for such calculations is stated. First, we have to consider a hypothesis based on data and then check the plausibility of the hypothesis. Given a hypothesized power law distribution from which our observed data is drawn, we are interested in finding that our hypothesis is a plausible one, given the data. We apply Kolmogorov-Smirnov (KS) statistic for estimating the *goodness of fit* for given data and calculate p-values based on KS statistic. If the p-values are larger than the level of significance, then we do not have sufficient proof to reject the hypothesis i.e., accept the null hypothesis otherwise, we reject the null hypothesis. Note that in this paper, we consider the level of significance as 0.05. To check this plausibility a code implemented earlier by Thelwall(2016c) for citation data has been used.

Table 3. Statistical values of Hooked Power law and lognormal distribution for altmetric data

	Hooked Power law				***Lognormal Distribution***			
	A	*B*	*ks*	*p value*	*μ*	*σ*	*ks*	*p value*
Twitter	2.235	6.709	0.032	0.481	1.701	1.525	0.040	0.266
Facebook	80.566	208.29	0.316	0.000	0.990	0.494	0.151	0.000
Blog	99.990	240.95	0.335	0.000	0.942	0.468	0.176	0.000
News	2.789	6.031	0.224	0.000	1.360	1.078	0.164	0.000

* (Based on software used in (Thelwall, 2016c)}

Discussion

The existence of power law characteristics have been widely explored for diverse data from many disciplinary domains, including citations, which are the most prevalent and useful impact indicator of the scholarly articles (Brzezinski, 2015; Peterson et al., 2010; Radicchi, Fortunato, & Castellano, 2008; Thelwall, 2016c, 2016a). Given that the altmetric mentions are found to correlate with citations (Thelwall & Nevill, 2018; Banshal et al., 2021; Shema et al., 2014; Thelwall, 2018) and that they are skewed in nature (Thelwall & Nevill, 2018; Thelwall & Wilson, 2016; Costas et al., 2016), it would be an interesting exercise to explore if altmetric mentions follow power laws. Yet, this property had not been put to the test so far. Our approach has been to investigate the plausibility of the existence of power law behaviours in altmetric mentions. The plots in the log-log space indicate existence of power laws in different kinds of altmetric mentions. It is also found that truncation of the data series can improve the closeness of fit. We obtained an increase in the R^2 parameter from about 1 to 9% by truncation of different altmetric data series such as Twitter, Facebook and the aggregate score Alt-score (Table 2). Subsequently, hooked power law and discretised lognormal behaviour have also been explored on bootstrapped samples (Table 3).

Power Law in Altmetrics

The first step in scrutinizing the existence of power laws in altmetric data is the log-log plot for mentions across several social media platforms- Facebook, Twitter, etc. and the composite score built from all individual scores. Altmetric.com also collects data from other platforms not mentioned here, which we have not included in the study due to the sparseness of data. The plot is shown in Figure 1, from which, it is observed that an approximate straight line can be fit to the data plotted on a log-log plot. This is the first characteristic of power laws and verifies the presence of power laws in altmetric mentions. These plots also specify the parameters of fit. These values help in analysing the nature of the plots and also determine where the majority of the distribution of data lies. The low (<=2) value of the power-law scaling exponent reflects an infinite (or divergent) mean, whereas mean is finite/converges for exponent>2. If $\alpha<2$, it means that most of the mentions lies in tail of the distribution. Further, the median exists for exponent (B) >1 and also, variance exists if exponent >3. The parameters of fit have been listed in Table 2.

In the table 2, value of R^2 indicates the variance explained by the model, high value of R^2 is necessary for acceptance of the power law distribution (Clauset et al., 2009). It is also a positive evidence for our first research question. Hence, we can draw the conclusion that altmetrics mentions do follow a power law. These parameters of fit provide evidence of existence and degree of fit, but do not prove that the power law is the sole model for the observed data.

Discretised lognormal and Hooked Power law in Altmetrics

After having conclusive evidence of power law in altmetric mentions, our second investigative point was to check the plausibility of discretised lognormal and hooked power law in altmetrics. To do so, we have drawn samples of 500 random observations from each category of data and then tried to estimate the population parameters for the same. Next, we applied KS statistic for estimating the *goodness of fit* for given data and calculated p-values based on KS statistic for each individual category. The results are summarized in Table 3.

Since the p-values of Twitter for both distributions is greater than the level of significance, it indicates that we don't have enough evidence to reject the hypothesis. Hence for twitter mentions, discretised lognormal and hooked power laws are also a plausible form of power law. Whereas in the case of Facebook, blog and news mentions, the p-values are less than the level of significance, which means these mentions do not provide a good fit for hooked and discretised lognormal distributions. Note that in this paper, we consider the level of significance as 0.05. There could be several factors for such low p-values. One could be that, these distributions are better observed only above a certain minimum value. Next could be that there are always some deviations because of the random nature of sampling.

In this approach also, altmetric mentions show similar trends like citations to some extent as both discretised lognormal and hooked power law are found to be plausible for the citation data (Thelwall, 2016c, 2016a). Though, the online mediums like blogs, news and Facebook do not show enough proof of the existence of discretised lognormal and hooked power law, Twitter is showing some indication towards the plausibility of these modified forms.

This discussion shows that altmetric data do follow power law distribution (Table 2). Truncation of some data in the tail of the distribution improves the degree of fit. Furthermore, discretised lognormal and hooked power law provide a good fit for Twitter (tweets). Other social media platforms like Facebook, blog and news mentions do not have a good fit for the same distributions. For these the power law provides the best fit. One could further explore other forms of power laws or exponential distributions, but it appears unnecessary as the degree of fit to power law is reasonably good as seen from Table 2 and data truncation.

Conclusion

We report here on the occurrence of power laws in altmetrics, where our data on altmetrics constitute counts of mentions and reads about scientific articles. While not entirely surprising, it is nevertheless interesting to see more and newer systems getting subsumed under the same law. Not only do the above categories follow power laws, but we find that the composite altmetrics index, created by combining individual altmetric indices, also follows a power law. Finally, we test some models, used by authors earlier in similar contexts, with the observed data we have gathered from Altmetric.com.

The study draws useful information on the nature of social media data of scholarly articles, chiefly whether they follow power law distributions which have been found to be ubiquitous in many natural and man-made systems. We find that the distribution of altmetric data does follow a power law, but with variable goodness of fit for the various social media data. For a closer fit, we tried using the discretised lognormal and hooked power laws. We find that to a fair degree of approximation ($R^2 > 0.9$) data from the four social media platforms and the composite data follow power laws. The degree of fit can be further improved by some data truncation in the tail. As future endeavours, one can address and discuss the implications of existence of the power law in Altmetrics.

Acknowledgments

The authors would like to acknowledge the support of Stacy Konkiel, Director of Research Relations at Digital Science, for providing access to Altmetric.com data.

References

Adamic, L. A. (2000). Zipf, power-laws, and pareto-a ranking tutorial. *Xerox Palo Alto Research Center, Palo Alto, CA, http://Ginger.Hpl.Hp.Com/Shl/Papers/Ranking/Ranking.html.*

Adamic, L. A., & Huberman, B. A. (2001). The Web's hidden order. *Communications of the ACM*, *44*(9), 55–60.

Banshal, S. K., Singh, V. K., & Muhuri, P. K. (2021). Can altmetric mentions predict later citations? A test of validity on data from ResearchGate and three social media platforms. In *Online Information Review: Vol. ahead-of-print* (Issue. *ahead-of-print*). https://doi.org/10.1108/OIR-11-2019-0364

Bouchaud, J.-P. (2001). Power laws in economics and finance: some ideas from physics. *Quantitative Finance*, *1*(1), 105–112. https://doi.org/10.1080/713665538

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, *137*, 85–86.

Brzezinski, M. (2015). Power laws in citation distributions: evidence from Scopus. *Scientometrics*, *103*(1), 213–228. https://doi.org/10.1007/s11192-014-1524-z

Chen, Y., & Leimkuhler, F. F. (1986). A relationship between Lotka's law, Bradford's law, and Zipf's law. *Journal of the American Society for Information Science*, *37*(5), 307–314.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, *66*(10), 2003–2019. https://doi.org/10.1002/asi.23309

Costas, R., Haustein, S., Zahedi, Z., & Larivière, V. (2016). Exploring paths for the normalization of altmetrics: Applying the Characteristic Scores and Scales. The 2016 Altmetrics Workshop, Bucharest, Romania. In The 2016 Altmetrics Workshop, Bucharest, Romania, 27 September 2016.

Duan, Y., & Xiong, Z. (2017). Download patterns of journal papers and their influencing factors. *Scientometrics*, *112*(3), 1761–1775. https://doi.org/10.1007/s11192-017-2456-1.

Egghe, L. (2005). Power laws in the information production process: Lotkaian informetrics. Elsevier Ltd.

Egghe, L. (2009). Performance and its relation with productivity in Lotkaian systems. Scientometrics, 81(2), 567-585.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, *13*(4). https://doi.org/10.2196/jmir.2012

Faloutsos, M., Faloutsos, P., & Faioutsos, C. (1999). On power-law relationships of the internet topology. *Computer Communication Review*, *29*(4), 251–261.

Gabaix, X. (2009). Power laws in economics and finance. *Annu. Rev. Econ.*, *1*(1), 255–294.

Gabaix, X. (2016). Power laws in economics: An introduction. *Journal of Economic Perspectives*, *30*(1), 185–206.

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology*, *65*(4), 656–669.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–323.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, *1*(2), 226–251. https://doi.org/10.1080/15427951.2004.10129088

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351. https://doi.org/10.1080/00107510500052444

Newman, M. E. J. (2007). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351. https://doi.org/10.1080/00107510500052444

Pareto, V. (1896). *Coursd'économie politique: professé à l'Universįté de Lausanne* (Vol. 1). F. Rouge.

Peoples, B. K., Midway, S. R., Sackett, D., Lynch, A., & Cooney, P. B. (2016). Twitter predicts citation rates of ecological research. *PLoS ONE*, *11*(11), 1–11. https://doi.org/10.1371/journal.pone.0166570

Peterson, G. J., Pressé, S., & Dill, K. A. (2010). Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(37), 16023–16027. https://doi.org/10.1073/pnas.1010757107

Price, D. J. D. S. (1965). Networks of scientific papers. *Science*, 510–515.

Price, D. J. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306.

Priem, J., & Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, *15*(7), 1–19.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *altmetrics: a manifesto*. 1–77. http://altmetrics.org/manifesto/

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272.

Radicchi, F., Fortunato, S., Castellano, C., & Moro, P. A. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(45), 17268–17272.

Redner, S. (1998). Rapid Note How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B*, *4*(2), 131–134.

Redner, S. (2005). Citation statistics from 110 years of Physical Review. *Physics Today*, *58*(6), 49–54. https://doi.org/10.1063/1.1996475

Rousseau, R. (1997). Sitations: an exploratory study. *Cybermetrics*, 1(1) available at: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Schubert, A., Glänzel, W., & Braun, T. (1987). Subject field characteristic citation scores and scales for assessing research performance. *Scientometrics*, 12(5-6), 267-291.

Shema, H., Bar-Ilan, J., & Thelwall, M. (2014). Do Blog Citations Correlate With a Higher Number of Future Citations? Research Blogs as a Potential Source for Alternative Metrics. *Journal of the Association for Information Science and Technology*, *65*(5), 1018–1027. https://doi.org/10.1002/asi

Thelwall, M. (2016a). Are the discretised lognormal and hooked power law distributions plausible for citation data? *Journal of Informetrics*, *10*(2), 454–470. https://doi.org/10.1016/j.joi.2016.03.001

Thelwall, M. (2016b). Are there too many uncited articles? Zero inflated variants of the discretised lognormal and hooked power law distributions. *Journal of Informetrics*, *10*(2), 622–633. https://doi.org/10.1016/j.joi.2016.04.014

Thelwall, M. (2016c). The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression. *Journal of Informetrics*, *10*(2), 336–346. https://doi.org/10.1016/j.joi.2015.12.007

Thelwall, M. (2018). Early Mendeley readers correlate with later citation counts. *Scientometrics*, *115*(3), 1231–1240. https://doi.org/10.1007/s11192-018-2715-9

Thelwall, M., & Nevill, T. (2018). Could scientists use Altmetric.com scores to predict longer term citation counts? *Journal of Informetrics*, *12*(1), 237–248.

Thelwall, M., & Sud, P. (2016). Mendeley Readership Counts: An Investigation of Temporal and Disciplinary Differences. *Journal of the Association for Information Science and Technology*, *67*(12), 3036–3050. https://doi.org/10.1002/asi

Thelwall, M., & Wilson, P. (2016). Mendeley Readership Altmetrics for Medical Articles: An Analysis of 45 Fields. *Journal of the Association for Information Science and Technology*, *67*(8), 1962–1972. https://doi.org/10.1002/asi

Wang, X., Fang, Z., & Sun, X. (2016). Usage patterns of scholarly articles on Web of Science: a study on Web of Science usage count. *Scientometrics*, *109*(2), 917–926. https://doi.org/10.1007/s11192-016-2093-0

Zipf, G. K. (1949). Human behaviour and the principle of least-effort. Cambridge MA edn. *Reading: Addison-Wesley*.

Data sources and their effects on the measurement of open access. Comparing Dimensions with the Web of Science

Isabel Basson[1], Marc-André Simard[2], Zoé Aubierge Ouangré[3], Cassidy R. Sugimoto[4] and Vincent Larivière[5]

[1] *isabel.basson@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal, Québec (Canada)
DST-NRF Centre of Excellence in Scientometrics and STI Policy; and Centre for Research on Evaluation, Science and Technology, Stellenbosch University, Stellenbosch (South Africa)

[2] *marc-andre.simard.1@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal, Québec (Canada)

[3] *zoe.aubierge.ouangre@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal, Québec (Canada)

[4] *sugimoto@indiana.edu*
School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, Indiana (USA)

[5] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal, Québec (Canada)

Abstract

With the amount of open access (OA) mandates at the funder and institution level growing, the accurate measurement of OA publishing is an important policy question. Existing studies have provided estimates of the prevalence of OA publications ranging from 27.9% to 53.7%, depending on the data source and period of investigation. This paper aims at providing a comparison of the proportion of OA publishing as represented in two bibliometric databases, Web of Science (WoS) and Dimensions, and assess how it affects the measurement of OA across different countries. Results show that publications indexed in Dimensions have a higher percentage of OA than those indexed by the WoS, especially for publications from outside North America and Europe. The paper concludes with a discussion of the cause and consequences of these differences, motivating for the use of more inclusive databases when examining OA, especially for publications beyond North America and Europe.

Introduction

Over the past 30 years, the democratization of the internet has allowed researchers to make their publications freely available on a worldwide scale. This practice, also known as open access (OA), allows anyone with an Internet connection to access, read, distribute, and download scientific publication for free with no legal or technical barriers. OA publishing is no longer a marginal phenomenon: for the most recent years, nearly half of all scholarly publications are estimated to be freely available online (Piwowar et al., 2018). This is due to a massive rise in OA mandates (Larivière & Sugimoto, 2018), the introduction of several new OA publishers and of OA options from legacy publishers, the creation of software that facilitate the production of publications (such as the Public Knowledge Project), and the rise of OA mega-journals such as PLOS ONE. The advantages of OA have been highlighted in the scientific literature such as a higher global visibility (Evans and Reimer, 2009), higher citation rates (Archambault et al., 2014; Piwowar et al.,2018), and a better use of taxpayers' money (Suber, 2003). Several studies have attempted to assess the overall share of OA publications in the scientific literature, with results ranging from 27.9% to 53.7%, depending on the data source and period of investigation (Piwowar et al., 2018; Archambault et al., 2014; European Commission, 2019). This study aims at providing a comparison of the proportion of OA as represented in two bibliometric databases, Web of Science (WoS) and Dimensions, and assess how the different coverage of these two databases may affect the measurement of OA across different countries.

Data sources

The Science Citation Index (SCI) was originally developed by Eugene Garfield (1955) to help librarians and researchers to find articles and journals relevant for their work through their references. Since it was impossible to manually cover the entire range of journals (~50 000 at the time; Garfield, 1955), only the most cited publications were indexed in the SCI and eventually in WoS. WoS was the main source of data used by researchers in the bibliometrics community for decades. However, over the past 15 years, there have been a multiplication of new bibliometric sources of data such as Scopus (2004), Google Scholar (2004), Microsoft Academic (2016), and more recently, Dimensions (2018), which all offer different coverage of the scientific literature. Given their different coverage and indexation practices, the use of a given bibliometric data source may affect the results obtained. Mongeon and Paul-Hus (2016) have shown that WoS' and Scopus' coverage differed considerably: WoS has a significantly lower coverage of research in social sciences and humanities from non-English-speaking countries compared to Scopus. Similarly, Dimensions has much broader coverage than both WoS and Scopus (Herzog, Hook, & Konkiel, 2020; Visser, van Eck, & Waltman, 2020), given that indexing is not based on selective criteria (e.g., citations) but rather on a technical component: the presence of a Digital Object Identifier (DOI). However, the effect of this larger inclusion on the measurement of uptake of OA remains largely unknown.

Open access and developing countries

Developing and developed countries have historically differed in terms of preferred approaches to OA. Developed countries tended to make use of repositories, with self-archiving mandates in place at many institutions (Jingfeng et al., 2012) and by funders (Sugimoto & Larivière, 2018). These mandates are supported by corresponding infrastructure, such as the government-funded PubMed repository or institutional repositories. These types of infrastructure are less prevalent in developing countries as reported by the Registry of Open Access Repositories (https:// http://roar.eprints.org/). On the other hand, authors from developing countries tend to make use of OA journals when rendering their publications OA (Sotudeh & Horri, 2008: 71). Several factors can explain the difference in self-archiving facilities and practices. For instance, developing countries do not necessarily have access to the same technological infrastructure (i.e., computer servers, computers, access to internet) as developed countries (Evans and Reimer, 2009). OA is built on the assumption that Internet access is a basic public utility that is available to everyone. This flawed assumption places developing countries at a significant disadvantage when discussing OA (Ouangré, 2020). For example, in 2018, nearly 75% of the African population did not have access to the internet (Broadband Commission, 2019). Various platforms have emerged aiming to increase online visibility of research publications from developing countries and to combat the effect of the serials crisis (Das, 2015), with the focus on supporting and launching OA journals. Such platforms include AJOL (Africa), AmeliCA (Latin America), and SciELO (Brazil).

Methods

We investigated all the journal-based publications indexed in WoS and Dimensions for the publication years 2015 to 2019 for which first author country affiliation data is available. We used Unpaywall's five-categories classification system (Piwowar et al., 2018) to identify the OA status of publications.

- **Gold:** Published in an OA journal that is indexed by the Directory of Open Access Journals (DOAJ).
- **Green only:** Toll-access on the publisher page but is free in an OA repository.
- **Hybrid:** Free under an OA license in a toll-access journal.
- **Bronze:** Free to read on the publisher page, but without a clearly identifiable license.
- **Closed:** All other publications, including those shared only on an Academic Social Network (ASN) or in Sci-Hub.

More generally, OA publications are defined in this paper as all publications that are not in the "closed" category. We also investigated publications by region. Each publication is assigned to a single country based on the country affiliation that appeared on the publications for the first author. Countries are then categorized by region. The classification system we used is the World Bank Country Classification by Region for 2020 (World Bank 2020).

Results

For the years 2015 to 2019 WoS indexes a total of 8,053,050 publications for which affiliation data is available, whereas Dimensions indexes 10,743,016 publications. Of these publications, 43.4% of WoS publications and 46.6% of the Dimensions publications are available as OA publications, as shown in Figure 1. The largest differences observed between the two datasets are for the "green only", and "bronze" categories, with a larger percentage of OA publications in WoS for the former, and a larger percentage in Dimensions for the latter. These global results are not surprising: as shown by Visser et al. (2020), most of WoS papers are actually indexed in Dimensions — although their metadata is often incomplete.

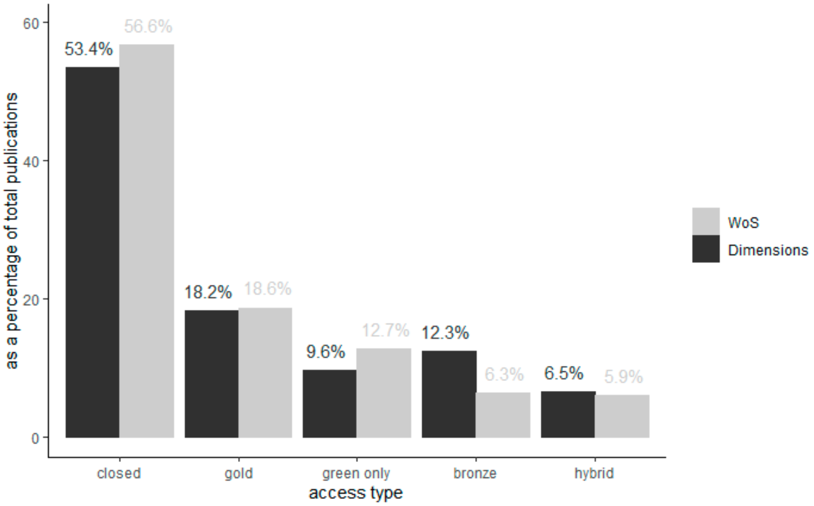


Figure 1 Percentage of open access publications, by access type and database, 2015-2019.

When we investigate publications by region (Figure 2) for the two databases we observe that the percentage of publications that are OA are similar to each other for North America, Europe and Central Asia. For the other regions, the percentage of OA publications in Dimensions are significantly higher than in WoS, especially for South Asia (+57.9%), Latin Americas and the Caribbean (+36.6%), the Middle East and North Africa (+33.5%) and, to a lesser extent, Sub-Saharan Africa (+12.4%).

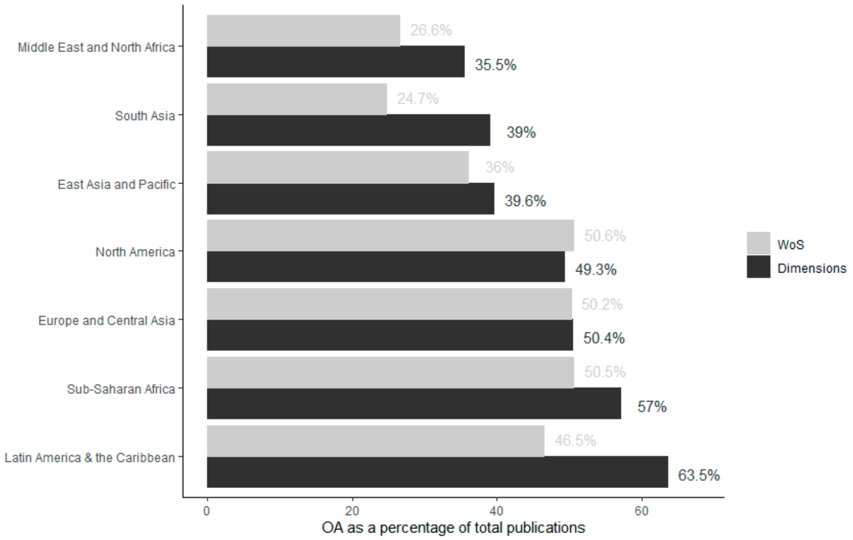


Figure 2 Percentage of open access publications, by region and database, 2015-2019.

To further examine these differences, we investigated the percentage of OA publications for the individual countries, with a specific focus on those countries for which a substantially higher percentage of publications are OA than in the other dataset. We only considered those countries for which more than 100 publications are indexed in both datasets for the period of investigation. In Figure 3 we present the results for the top 10 countries with the largest difference between the two datasets. When examining the top 10 countries with the largest difference in favour of Dimensions we can observe substantial differences in percentage of OA publications, ranging from 24.4% to 46.0%. For WoS, we see a smaller range, from 2.4% to 9.5%.

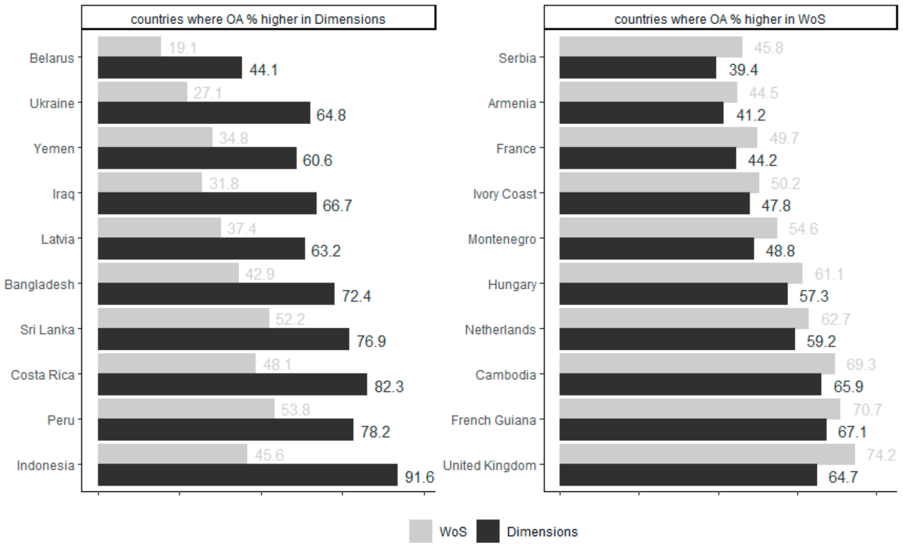


Figure 3 Percentage of open access publications, for selected countries and database, 2015-2019.

Discussion

Our results show that the measurement of OA differs when using WoS or Dimensions, and that the difference is more striking for authors from outside North America, and Europe and Central Asia. Given the Western bias of journals indexed in WoS (Mongeon and Paul-Hus, 2016)—which are also indexed in Dimensions (Visser et al., 2020)—the measurement of OA in these regions does not vary much in the two databases. However, for the other regions, which generally have their scientific production underestimated in WoS, the additional publications that are indexed in Dimensions are much more likely to be OA. More specifically, as Dimensions has much broader indexing, this higher percentage of OA publications is potentially due to the inclusion of smaller national journals. Further research is required to investigate the characteristics of the additional publications included in Dimensions.

This has further implications for the distribution of the different types of OA, as the literature suggests that the countries generally represented in WoS are also those that tend to more often make use of self-archiving (green OA). This potentially explains the larger percentage of green OA publications in WoS, as various mandates are applicable, and many repositories are available, to these authors. The higher percentage of bronze OA in Dimensions could reflect the inclusion of many non-DOAJ listed journal publications in Dimensions. It is likely that these bronze OA publications are in journals not published by the major publishers and lack the same level of standardization in metadata, resulting in difficulty classifying the publications and their inability to be indexed in DOAJ.

Lastly, this difference in measurement is most clearly illustrated at the level of countries. If WoS is used to measure OA for countries, it can significantly underestimate the percentage of OA publications for some countries in comparison to a more inclusive database such as Dimensions. Just as OA aims to provide visibility and access to research publications beyond toll-access journals, Dimensions provides a lens to investigate a broader number of publications, as opposed to only those that are considered to be the most relevant or core by WoS. What our analysis has shown is that the measurement of OA may differ significantly when one looks beyond this scientific core. Ultimately, Dimensions has the potential to be a more suitable platform for a more inclusive measurement of OA uptake, especially of publications by authors from outside North America, Europe and Central Asia, where mandates, repositories or infrastructures are not as prevalent.

Acknowledgments

This work is based on research funded by the International Development Research Centre (IDRC) Science Granting Councils Initiative in Sub-Saharan Africa program under project 109272.

References

Archambault, É., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebout, L., & Roberge, G. (2014). *Proportion of open access papers published in peer-reviewed journals at the European and world levels—1996–2013*. European Commission. Retrieved from: http://science-metrix.com/en/publications/reports/proportion-of-open-access-papers-published-in-peer-reviewed-journals-at-the.

Broadband Commission. (2019). *The State of Broadband 2019: Broadband as a Foundation for Sustainable Development*. ITU/UNESCO Broadband Commission for Sustainable Development. Retrieved from: https://www.itu.int/dms_pub/itu-s/opb/pol/S-POL-BROADBAND.20-2019-PDF-E.pdf.

Das, Anup Kumar (2015). *Serials Crisis*. In: Mishra, S. & Satija, M.P. (eds.), *Open Access for Researchers, Module 1: Scholarly Communication*. Paris: UNESCO, pp. 44-67. ISBN 9789231000782.

European Commission. (2019). *Trends for open access to publications*. Retrieved from: https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/trends-open-access-publications_en.

Evans, J.A. & Reimer, J. (2009). Open access and global participation in Science. *Science*, 323(5917), 1025

Garfield, E. (1955). Citation indexes for science. *Science*, 122(3159), 108-111.

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, 1(1), 387-395.

Jingfeng, X., Gilchrist, S.B., Smith, N.X.P, Kingery, J.A., Radecki, J.R., Wilhelm, M.L., MArrison, K.C., Ashby, M.L. & Mahn, A.J. (2012). *A review of open access self-archiving mandate polices*. Portal: Libraries and the Academy, 12(1), 85-102.

Larivière, V., & Sugimoto, C. R. (2018). Do authors comply when funders enforce open access to research? *Nature*, 562(7728), 483-486.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106(1), 213-228.

Ouangré, Z. A. (2020). *Le comportement dans la recherche d'information des étudiants au doctorat en médecine au Burkina Faso*. Doctoral thesis, Université de Montréal. Available at : https://papyrus.bib.umontreal.ca/xmlui/handle/1866/23397

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A, West, J., & Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, 6, e4375.

Sotudeh, H. & Horri, A. (2008). Great expectations: the role of open access in improving countries' recognition. *Scientometrics*. 76(1):69–93. DOI: 10.1007/s11192-007-1890-x.

Suber, P. (2003). *The taxpayer argument for open access*. SPARC Open Access Newsletter. 65. Retrieved from: https://dash.harvard.edu/bitstream/handle/1/4725013/suber_taxpayer.htm.

Visser, M., van Eck, N. J., & Waltman, L. (2020). *Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic*. arXiv preprint arXiv:2005.10732.

World Bank. (2020). *World Bank Country and Lending Groups.* Retrieved from: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

Journal flipping to Open Access: The perception of Spanish journal managers

Núria Bautista-Puig[1], Carmen López-Illescas[2] and Henk F. Moed[3]

[1]*Nuria.Bautista.Puig@hig.se*
University of Gävle, Department of Industrial Management, Industrial Design and Mechanical Engineering, Kungsbäcksvägen 47, 801 76 Gävle, Sweden

[2] *carmlopz@gmail.com*
University Complutense of Madrid, Information Science Faculty, Department of Information and Library Science, SCImago Group, Spain

[3] *henk.moed@uniroma1.it*
Sapienza University of Rome, Italy

Abstract

There is a growing interest in determining the factors that influence a journal's flipping to Open Access (OA). Using semi-structured interviews combined with bibliometric indicators, this paper uncovers the perception of Spanish managers related to Open Access and the decision to flip. The key research questions are twofold: How well do bibliometric measures reflect the changes in the status of the journal? How do journal managers perceive the flipping process? In order to answer these, twelve semi-structured interviews were conducted with journal managers of Spanish Journals. The findings suggest the great majority of managers are aware of the indicators, but only two mentioned they reflect their reality. An increase in the number of submissions, visibility, or internationalization since the transition is perceived as a benefit while the loss of interchanges with other institutions is seen as the major drawback. Although flipping to OA is perceived by the managers to have many advantages, it raises some challenges too, especially the need for funding, lack of resources capacity for technical support, and the creation of alliances.

Introduction

Several studies have analyzed the motivations to publish in Open Access (OA) journals at different levels. At the author level, in a survey comparing OA-authors and no-OA authors, Swan and Brown (2004) found that the main motivations for publishing in OA, as opposed to publishing in subscription-based journals, were free access, faster publication times, and larger readerships. In a large-scale international survey of authors' perceptions, Rowlands, Nicholas, and Huntingdon (2004) found that reputation, impact factor, or speed of refereeing are essential in authors' choosing OA. Based on surveys and interviews, Kim (2010) investigated the factors that motivate or impede faculty participation in self-archiving practices, ranging from web pages to OA repositories. He identified seven factors (in descending order of importance): (a) altruism (the idea to provide OA benefits to users); (b) perceived self-archiving culture; (c) copyright concerns; (d) technical skills; (e) age; (f) perception of a non-harmful impact of self-archiving on tenure and promotion; and (g) concerns about additional time and effort. A longitudinal analysis by Xia (2010) analyzed the changing pattern of scholars' attitudes OA journal publishing from 1991 to 2018. This author observed an increase in the publication and awareness rates for OA journals despite the concern related to the prestige of the journal. In a similar study, Togia and Korobili (2014) explored the attitudes and perceptions toward OA presented in different studies and found differences across countries and disciplines. Free access which facilitates wider dissemination is a strong incentive, while the author-pays model, the quality of peer-review, and the impact of the journals are perceived as major concerns. Later, Nelson and Eggett (2017) surveyed chemistry authors from OA journals to determine why they chose hybrid journals. Among the main reasons, funding mandates and receiving higher numbers of citations are highlighted. Furthermore, altruistic reasons, such as providing scientific results to the wider public and to other researchers who might not have the financial means to obtain the articles otherwise, were found to play an important role. Rowley et al.

(2017) revealed a high level of uncertainty about future intentions in scholars' attitudes and behaviors toward publishing their research in OA journals, with small differences between disciplines'. In a recent study, Heaton, Burns, and Thoms (2019) surveyed 250 authors at Utah State University about their motivation of publishing OA articles. In this study, the ability to pay publication charges, disciplinary colleagues' positive attitudes toward OA, and personal feelings such as altruism and desire to reach a wide audience were mentioned as main factors.

Following up on a similar 2013 survey, a study published in 2021 tracked the changes in attitude towards OA publishing among faculty in Information Studies schools (Peekhaus, 2021). In agreement with recent research, this study revealed that engagement with OA has increased significantly between 2013 and 2018, but it observed a high level of uncertainty about future perception of OA. At the same time, the work outlined that some of the historical concerns of OA publishing are dissipating, such as those regarding the perception of quality, and underlined that even the willingness to pay for articles processing charges (APC) has increased significantly.

While there are many studies that explore OA perceptions at the *author* level, few studies addressed the perception of journal *managers*. From the perspective of the senior managers, Wakeling et al. (2019) examined the motivations for launching an OA mega journal (OAMJ). Two motivations were in line with OA: a) supporting the OA movement, which can be seen as a societal benefit; and b) the 'effect change' which is perceived as an opportunity for experimentation (e.g. open peer review) or a change at the systemic level (e.g. transform scholarly communication). In the Spanish context, Segado-Boj, Martín Quevedo and Prieto-Gutiérrez (2018) conducted 15 interviews with managers of Spanish journals to study their perception towards open access, open peer review, and altmetrics. Open access is perceived as a positive factor as managers highlighted many advantages (e.g., improving scientific dialogue, availability of the content, achieving higher visibility). However, this model also generated concerns: the difficulty of making the magazine profitable or the creation of an economic barrier for authors that could not afford the payment. Robinson and Scherlen (2009) surveyed managers of dozens of journals in criminology and criminal justice. These authors found that the managers were highly supportive of the OA principles, however, they had a more favorable view of traditional journals rather than of OA journals.

At the Spanish level, a few studies analyzed the perceptions of OA authors in Spain. Hernández-Borges et al. (2006) surveyed medical authors to determine their awareness and attitude towards OA publishing and the ''author pays'' model. These authors were found to have a low level of awareness of this model as well as a low acceptance of journals charging author fees. Ruiz-Pérez & Delgado-López-Cózar (2017) surveyed 554 researchers from different fields and concluded that 76% consider OA beneficial for their discipline. Serrano-Vicente, Melero, & Abadal (2016) analyzed the awareness of open access among the academic staff of a Spanish university and found that many respondents supported OA. In addition, the decision to publish in OA is directly related to academic reward and professional recognition. However, to the best of our knowledge, no previous study has conducted an analysis at the manager level in the Spanish context.

Methods

The use of qualitative interviews has the potential to detect issues not covered in the literature. Interviewing journals can provide answers to the following questions: How well do bibliometric measures reflect the changes in the status of the journal? How do journal managers perceive the flipping process?

Twelve semi-structured interviews (see interview guide at Bautista-Puig et al., 2021) were conducted with managers from Spanish journals collected in the first study by the current authors (Bautista-Puig et al., 2020). For more information the reader is referred to this article.

This technique provides a flexible tool for small-scale research as questions can be adapted, allowing for unexpected answers whilst also clarifying the interviewee's response. It also allows for a deeper insight into the respondent's views. The sampling procedure identified the participants through contact information with the journal websites. Out of 24 journals identified in the first round, twelve responded positively to the request to participate, two declined and ten did not respond after sending two reminders. All the interviews were conducted online, digitally recorded, and transcribed for analysis. The length of the interviews ranged from 40 to 90 minutes. The answers were coded subjected to a thematic analysis (Braun and Clarke, 2006) in the process to facilitate further analysis. These codes represent themes which emerged by combining insights from existing literature with the data. A data matrix was created with variables identifying the questions and values identifying specific answers to a question. Secondary data analysis of bibliometric indicators of each journal from SCImago Journal Rank (SJR), a database with journal indicators based on data from Elsevier's Scopus, complemented the primary data, highlighting the importance of adopting a mixed-method approach. Adoption of these methods enabled the current authors to observe how open access is conceptualized and operationalized across different journals from different fields, and to discuss the motivations underpinning Open Access. Table 1 lists the participants by subject category associated with the journal, the age of the journal, and the validated transition year. The sample represents different disciplines from philosophy to engineering. In total, 12 journals were interviewed, between June 2020 and October 2020.

Table 1. Participants characteristics

Journals	*Branch of science (based on the LCC Subject Category)*	*Starting year and age of the journal*	*OA year validated*
Participant 1	History (General) and history of Europe: History (General)	1964 (53 years)	2006
Participant 2	Geography. Anthropology. Recreation: Anthropology: Ethnology	1944 (73 years)	2006
Participant 3	Medicine: Surgery	1975 (42 years)	2006
Participant 4	Geography. Anthropology. Recreation: Geography (General)	1940 (77 years)	2006
Participant 5	Science: Geology	1966 (51 years)	2003
Participant 6	Technology: Home economics: Nutrition. Foods and food supply	1960 (57 years)	2006
Participant 7	Science: Botany	1879 (38 years)	2006
Participant 8	Technology: Electrical engineering. Electronics.	1949 (68 years)	2006
Participant 9	Science: Biology (General): Ecology	1945 (72 years)	2006
Participant 10	Technology: Home economics: Nutrition. Foods and food supply |	2012 (5 years)	2013
Participant 11	Language and Literature: Philology. Linguistics | Philosophy	1941 (76 years)	2006
Participant 12	Philosophy. Psychology. Religion: Philosophy (General)	1952 (65 years)	2009

The objective of the interviews is a) to gain a better understanding of the journal flipping process by journal managers' view, and b) to determine the contextual factors that influence the (un)success of OA flipping journals.

Results

Performance of bibliometric indicators

Half of the journal representatives (Nos. 3, 5, 6, 7, 9, 10, 12) were familiar with bibliometric information while the indicator showing the countries citing a journal surprised them the most. Even though many managers see the use of journal-level indicators as part of their editorial duty to try to improve the indicators (Krell, 2010), the lack of indicators awareness was associated with a lack of time or staff for monitoring this information (Nos. 1, 2, 3). In terms of how well bibliometric measures reflect the development of the journal, only two managers agreed they 'reflect the reality' (but some have concerns in only attributing this change to OA).

Benefits and cons of the transition and barriers identified

Table 2 summarizes the benefits and disadvantages related to the transition towards OA identified by the journal's interviewee ranked by descending order of frequency. The number of benefits outstands the number of cons, denoting a positive attitude towards the benefits perceived by the journals since the flipping. Within the main benefits, the increase in visibility is highlighted. This is in line with findings from several previous studies at the author level (Warlick and Vaughan, 2007; Williams et al., 2019) and at the journal level (Segado-Boj, Martín Quevedo, and Prieto-Gutiérrez, 2018). Almost half of respondents observed an increase in the number of submissions and publications. However, as pointed out by Participant 10, this increase in number of submissions does not imply an increase in the 'quality', and the selection of papers even becomes more difficult. In contrast, many participants did not identify any drawbacks (e.g. nrs. 1, 2, 3, 5 or 6). The loss of interchanges was pointed out by five of the interviewees (1, 2, 3, 8, 9). The lack of resource capacity was also mentioned by some participants (1, 8, 10, 12), meaning that an increase in the number of received manuscripts also leads to an increase in work. Such an increase could become unmanageable, and it could become especially difficult for journals managed by few people (1, 8, 10, 12). Participant 8 even considered the increase in the number of submissions as negative for them, stating 'we didn't have the capacity to manage this large number, we need to find enough reviewers to follow the publication standards'.

Table 2. Benefits and drawbacks identified by the interviewees (participant number between brackets)

Benefits	***Cons***
Increase in visibility (1, 2, 3, 6, 7, 9, 10, 11)	Loss of interchanges between institutions (1, 2, 3, 8, 9)
Increase in the submissions and publications (1, 2, 3, 6, 9, 10)	Lack of resource capacity: increase of workload (1, 8, 10, 12) or lack of knowledge about the processes (e.g. request a DOI, obtain data in xml…) (10)
Increase of the internationalization (2, 3, 9)	Competence with big publishers (4, 5,10, 11)
Chance to publish in other languages (1, 2, 9)	Loss of national languages i.e. more papers on English-speaking languages (12) and loss of readers (2)
Better accessibility (1, 3)	No increase in the submissions (12)
Reductions of costs e.g. images in a paper (5, 9)	
Acceptance and happiness with the decision within the editorial board (7) or the authors (12)	
The prestige of the host institution and the journal (4)	
Increase of the citations/JIF (6)	
An increase of the acceptance of peer review requests (as the journal is more widely recognized) (10)	
Adoption of a platform for the management of the journal (12) i.e. Open Journal System (OJS)	
Better quality of the papers (9)	
Having the capacity to take decisions about the management of the journal (10)	

Finally, it is also important to observe that some interviewees attributed the evolution of the journal not to OA, but rather to other factors. As an example, the integration of the journal into Journal Citation Reports (JCR) and the change of the journal's publication language to English were highlighted as the main drivers of change (nrs. 7, 8). This perception was also shared by Participant 10, who pointed out that other changes took place in his journal at the same time (namely its entry into SCIELO, a large Brazilian literature database) and, therefore, concluded it was not possible to draw a firm conclusion about the benefits directly related to OA. However, the decision to enter into Scielo was partly driven by OA. Another reason is the interdisciplinarity of the journal (1, 2). For Participant 5, the change was described in terms of the transition from paper-based to electronic publishing, not as a flip to OA itself.

Challenges for OA in Spain

Regarding the question of the main challenges towards OA in general and particularly in Spain, different views were expressed (Table 3). The need for funding for OA is a widely shared challenge. Despite the fact that the majority of journals receive funding from the host institution, the problem is the scarceness of resources which allow only for subsistence. Participant 10 believed there must be an institutional initiative to maintain open access without charging authors and readers. In addition to money, the lack of recognition was also pointed out by the journal managers. They feel they do not have any additional recognition with incentives or rewards for being OA (7). Another widely accepted challenge mentioned by the interviewees is the lack of resources capacity or technical support in general. Participant 2 commented that it was difficult to differentiate the roles within the journals between the management, the authors, and the technical staff because they are all sharing tasks. For instance, technical staff is not only dedicated to one single journal but to five more journals, and the managers are performing management tasks of the journal. In social sciences and humanities, technical staff is even scarcer.

Table 3. Challenges identified by the interviewees

Challenges identified
Lack of resources capacity or technical support (1, 2, 8, 10, 12)
The need for funding (1, 7, 10, 11)
Get institutional help or agreements to maintain OA without charging authors and readers (5, 6, 10)
Competence with the big publishers, especially when you are a 'small journal' (7, 10, 12)
'Appreciation of OA' and change of mind of researchers (3).
Maintain and value national publication languages (4)
Resistance to flip to OA, especially in Humanities (2)

Conclusions

Although some studies analyzed the motivations behind journal flipping, limited research has been conducted on how this transition is perceived by journal managers. The current paper provides insights into the elements, benefits, and challenges of OA flipping from a manager perspective. This research adds to the growing body of literature on Open Access, accentuating the complexity involved in incorporating OA in journals in the Spanish context. As regards the question how bibliometric measures reflect the changes in the status of the journal, findings suggest that the great majority of journals are aware of the indicators, but only two mentioned they reflect their reality since the transition. The results also show that the flipping process was largely perceived as beneficial. The number of benefits outstands the number of cons, denoting a positive attitude towards the journal's performance since the flipping. Among the main benefits, the increase in visibility or number of submissions is highlighted, while the loss of interchanges and the resource capacity are seen as major drawbacks.

This study has several limitations. First, this study is based on a small sample. As such, results cannot be generalized for the whole population and only subjective conclusions can be generated. Second, the great majority of journals that accepted to make the interview, belong to the same host institution and were subjected to the same set of requirements. This can affect the independence of the responses provided and can limit the ability of the findings to be generalized to other contexts. Third, the extent to which the perceptions on OA flipping of managers of flipping journals differ from those of non-OA journals has not been addressed in the current study and could be examined in a follow-up study. In addition, this research has highlighted a number of issues that require further investigation. The related literature has shown that the positive attitude toward OA among journal managers is increasing. We envision further study in this area through surveys and interviews with more journal managers in various fields which can analyze both the short and long-term effects on the journal. Contributions from other countries would complement this study and increase its generalizability.

References

Bautista-Puig, N., López-Illescas, C., de Moya-Anegon, F., Guerrero-Bote, V., & Moed, H. F. (2020). Do journals flipping to gold open access show an OA citation or publication advantage? Scientometrics, 124(3), 2551-2575. https://doi.org/10.1007/s11192-021-03939-6

Bautista-Puig, N., López-Illescas, C., & Moed, Henk F. (2021). Interview Guide, Open Access Study. figshare. Dataset. https://doi.org/10.6084/m9.figshare.14473866.v1

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Heaton, R., Burns, D., & Thoms, B. (2019). Altruism or self-interest? Exploring the motivations of open access authors. *College and Research Libraries*, *80*(4), 485–507.

Hernández-Borges, A. A., Cabrera-Rodríguez, R., Montesdeoca-Melián, A., Martínez-Pineda, B., De Arcaya, M. L. T. Á., & Jiménez-Sosa, A. (2006). Awareness and attitude of Spanish medical authors to open access publishing and the "author pays" model. *Journal of the Medical Library Association*, *94*(4), 449–451.

Kim, J. (2010). Faculty self-archiving: Motivations and barriers. *Journal of the American Society for Information Science and Technology*, *61*(9), 1909-1922.

Krell, F. T. (2010). Should editors influence journal impact factors? Learned Publishing, 23(1), 59–62. https://doi.org/10.1087/20100110

Nelson, G. M., & Eggett, D. L. (2017). Citations, mandates, and money: Author motivations to publish in chemistry hybrid open access journals. *Journal of the Association for Information Science and Technology*, *68*(10), 2501–2510. https://doi.org/10.1002/asi.23897

Peekhaus, W (2021). A cohort study of how faculty in LIS schools perceive and engage with open-access publishing. *Journal of Information Science,* 47(1), 16–28.

Robinson, M., & Scherlen, A. (2009). Publishing in Criminology and Criminal Justice: Assessing Journal Editors' Awareness and Acceptance of Open Access. *International Journal of Criminal Justice Sciences*, 4(2), 98.

Rowlands, I., & Nicholas, D. (2004). Their Last Published Paper. *Learned Publishing*, *17*(4), 261–273.

Rowley, J., Johnson, F., Sbaffi, L., Frass, W., & Devine, E. (2017). Academics' behaviors and attitudes towards open access publishing in scholarly journals. *Journal of the Association for Information Science and Technology*, *68*(5), 1201-1211.

Ruiz-Pérez, S., & Delgado-López-Cózar, E. (2017). Spanish researchers' opinions, attitudes and practices towards open access publishing. *Profesional de la Información*, *26*(4), 722-734.

Segado-Boj, F., Martín Quevedo, J., & Prieto-Gutiérrez, J. J. (2018). *Percepción de las revistas científicas españolas hacia el acceso abierto, open peer review y altmetrics*. *Ibersid: revista de sistemas de información y documentación* (ISSNe 2174-081X; ISSN 1888-0967), 12(1), 27-32.

Serrano-Vicente, R., Melero, R., & Abadal, E. (2016). Open access awareness and perceptions in an institutional landscape. *The journal of academic librarianship*, *42*(5), 595-603.

Swan, A., & Brown, S. (2004). Authors and open access publishing. Learned Publishing, 17(3), 219–224. https://doi.org/10.1087/095315104323159649

Togia, A., & Korobili, S. (2014). Attitudes towards open access: A meta-synthesis of the empirical literature. *Information Services & Use*, *34*(3-4), 221-231.

Wakeling, S., Creaser, C., Pinfield, S., Fry, J., Spezi, V., Willett, P., & Paramita, M. (2019). Motivations, understandings, and experiences of open-access mega-journal authors: Results of a large-scale survey. *Journal of the Association for Information Science and Technology*, *70*(7), 754–768. https://doi.org/10.1002/asi.24154

Williams, S. C., Farrell, S. L., Kerby, E. E., & Kocher, M. (2019). Agricultural researchers' attitudes toward open access and data sharing. *Issues in Science and Technology Librarianship*, (91).

Xia, J. (2010). A longitudinal study of scholars attitudes and behaviors toward open-access journal publishing. *Journal of the American Society for Information Science and Technology*, *61*(3), 615-624.

The Place of Preprint Citations in the IMRaD Structure: a Study of PLOS Journals

Marc Bertin[1] and Iana Atanassova[2]

[1] *marc.bertin@univ-lyon1.fr*
Laboratoire ELICO, Université Claude Bernard Lyon 1, France

[2] *iana.atanassova@univ-fcomte.fr*
CRIT, Université de Bourgogne Franche-Comté, France

Abstract

The role of preprints in the scientific production and their part in citations have been growing over the past 10 years. In this paper we propose to study the citations of preprints in several different aspects: the progression of these citations over time, the sections of the articles in which they are most likely to appear and the percentage of such citations with respect to all citations in articles. We have processed the PLOS dataset that covers 7 journals and a total of about 240000 articles up to January 2021. The results show that preprint citations are found with the highest frequency in the Method section of articles, though small variations exist with respect to journals. The PLOS Computational Biology journal stand out as it contains more than three times more preprint citations than any other PLOS journal. The relative parts of the different preprint databases are also examined. While ArXiv and bioRxiv are the most frequent citation sources, bioRxiv's disciplinary nature can be observed as it is the source of more than 70% of preprint citations in PLOS Biology, PLOS Genetics and PLOS Pathogens.

Introduction

In recent years, the growing role of preprints in the publication process and the creation of a large number of publicly available preprint databases has led to an increasing interest by the research community in scientometrics (see (Berg 2017; Kaiser 2017; da Silva 2017a,b,c)). Indeed, a growing number of online repositories provide access to new research that is still in production and not yet validated by peers. This contributes to the acceleration of the dissemination of science (Larivière et al., 2014), (Abdill & Blekhman, 2019). The arrival of bioRxiv on the ArXiv model and its operation (see the work of Pinfield (2001) who is interested in the use of this type of service by physicists) is presented by (Berg et al., 2016). (Fu & Hughey, 2019) show that having a paper in bioArxiv is correlated with a high altmetric value.
Beyond dissemination of recent results, preprints are also present in publications as cited references. (Frasr et al., 2020) show that papers submitted to bioRxiv receive more attention from the scientific communitythan other pubications. This implies that in the publication process, the citation of work from preprint databases is becoming now a part of the publication cycle (Hoy, 2020; Penfold & Polka, 2020). (Desjardins-Proulx et al., 2013) question the relationship between the preprint databases and the publishers with regard to the quality, but also the relevance of this type of production. Other ethical considerations arise such as plagiarism (Giles, 2003).
Direct citations to preprint databases such as arXiv, RePEc, SSRN and PMC all increased steadily from 2000 to 2013 according to the Scopus study by (Li,Thelwall & Kousha, 2015). Subsequent health crises mobilize research efforts and the role of preprints as an under-utilized mechanism for accelerating the dissemination of scientific findings is discussed by (Johansson et al., 2018) and (Majumder & Mandl, 2020). More broadly, this phenomenon has invited community researchers to produce datasets to include the production of arXiv in a recommendation system (Saier & Färber, 2019).

Traditionally, the knowledge built through scientific articles relies on originality and refers to previous and peer-reviewed work, thus participating in the cumulative structure of scientific knowledge through the act of citations. In fact, (Fry, Marshall & Mellins-Cohen, 2019) that the purposes of preprints are diverse: from increased dissemination speeds for authors to an invitation to critical reading for work that has not been peer-reviewed by the reader. The question that arises today is the place of preprints alongside scientific articles and their citations: how do preprints contribute to the construction of new scientific knowledge while their validity, as sources, is not established by peer review. Will the use of preprint citations reinforce the critical attitude expressed in articles or will it weaken the edifice of knowledge? To answer this question, one of the necessary steps is to study the nature and place of preprints in scientific journals.

In this paper we propose to study the citations of preprints with respect to their position in the different journals and the IMRaD (Introduction, Method, Result and Discussion) structure of articles. By processing the PLOS dataset that contains 7 journals mainly in the bio-medical domain, our study aims to provide evidence to answer the following questions:

- What is the part of the progression of preprint citations in articles over the last 10 years?
- In which sections of the rhetorical structure of articles are preprint citations most likely to appear?
- How do preprint citations relate to the overall number of citations in articles and in the different sections of the IMRaD structure.

As preprints contain recent results and methods that have not yet been validated by peers, one important question is whether these methods and results are actually cited and used in peer-reviewed journals. Knowing whether the preprints belong to one section rather than another is an indicator of the role of preprints in the construction of knowledge.

Method

In this study, we investigate the place of preprint citations within the IMRaD (Introduction, Method, Result and Discussion) structure of articles.

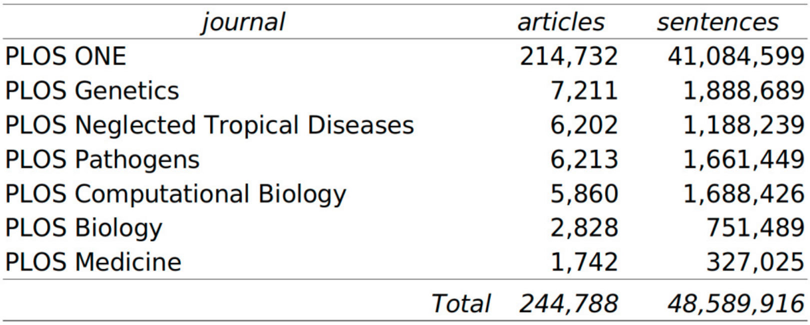
Dataset

We have used as a dataset all research articles published by PLOS up to January 2021. This dataset contains seven different journals and the majority of the articles is in the biomedical domain. The largest journal, PLOS ONE, is multidisciplinary and contains a large number of articles in various other domains. Table 1 gives the overall size of the dataset and the number of articles and sentences per journal.

The editorial requirements of PLOS journals impose using the IMRaD structure for research articles and the vast majority of the articles follow this pattern. This means that the four main section types (Introduction, Method, Result, Discussion) are present in all articles, while the order of these sections can vary according to the discipline and the journal.

In our experiment, we have considered only research articles that follow considered only research articles that follow this IMRaD pattern.

Table 1. PLOS dataset (up to January 2021)



journal	*articles*	*sentences*
PLOS ONE	214,732	41,084,599
PLOS Genetics	7,211	1,888,689
PLOS Neglected Tropical Diseases	6,202	1,188,239
PLOS Pathogens	6,213	1,661,449
PLOS Computational Biology	5,860	1,688,426
PLOS Biology	2,828	751,489
PLOS Medicine	1,742	327,025
Total	*244,788*	*48,589,916*

Processing pipeline

The dataset in available in XML format following the JATS (Journal Article Tag Suite) that is specifically designed to facilitate the processing of scientific papers. Our processing pipeline contains the following main steps:

1. Extraction of article metadata and bibliography items
2. Extraction of section titles and section classification
3. Processing of section content: paragraph and sentence segmentation
4. Processing of citations and linking them to bibliography items
5. Identification of preprint citations, their positions in the sections and all metadata related to the cited items.

Section classification

The first non-trivial processing step is the classification of section into the four section types ("Introduction", "Method", "Result" and "Discussion").

About 57% of the sections in the dataset are labelled as "intro", "method", "results" etc. in the "sec-type" attribute of the XML section element. We used these labels where possible. The remaining sections were classified using their section titles. Sets of regular expressions were manually designed to account for the possible variations in the titles. For example, the Discussion section was found under 488 different titles such as: 'Discussions', 'Discussion on practical implementation of proposed handover strategy', 'Summary and Discussion', 'Discussion and implications', 'General Discussion', 'Strengths and Limitations of the Study', 'Discussion/Conclusion', etc.

Some of the articles contain sections with titles that are subject-specific and could not be classified in this way. Such articles were left out. Among the total of 263,542 articles in the dataset, 234,592 (89%) articles contain the four section types and were used for our study.

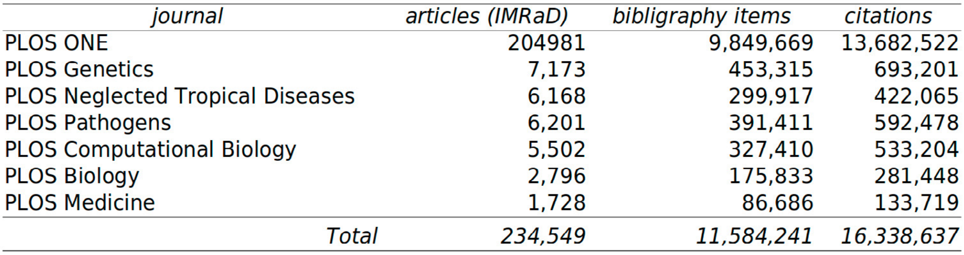
Processing of Citations

Article content was segmented into paragraphs and sentences. Citations were identified and linked to bibliography items using the existing annotations in the XML source files (i.e. xref elements that point to bibliography items).

The presence of citation ranges (e.g. [7]-[11]) requires supplementary processing as each range in represented in the XML structure of the document by only two xref elements that are linked to bibliography items. The remaining citations (e.g. [8], [9] and [10] in the range [7]-[11]) do not have xref elements in the XML. The creation of these new links accounts for about 9.5% of all citations in the dataset.

Table 2 presents the number of bibliography items and citations per journal.

Table 2. PLOS dataset: IMRaD articles, bibliography items and citations per journal



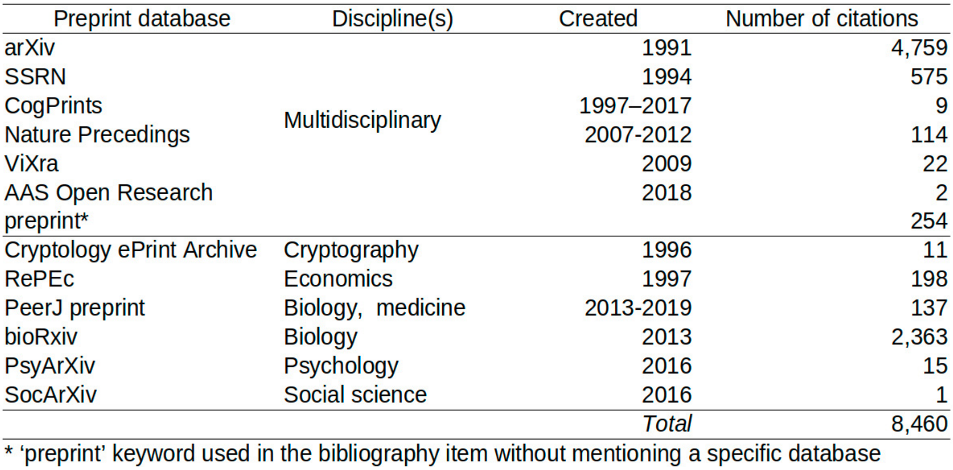
journal	*articles (IMRaD)*	*bibligraphy items*	*citations*
PLOS ONE	204981	9,849,669	13,682,522
PLOS Genetics	7,173	453,315	693,201
PLOS Neglected Tropical Diseases	6,168	299,917	422,065
PLOS Pathogens	6,201	391,411	592,478
PLOS Computational Biology	5,502	327,410	533,204
PLOS Biology	2,796	175,833	281,448
PLOS Medicine	1,728	86,686	133,719
Total	*234,549*	*11,584,241*	*16,338,637*

Identification of Preprint Citations

In order to account for all possible preprint databases, we have made a list of repositories from several sources[1]. Preprint citations were identified by regular expressions that were matched against bibliography items metadata. From the entire list of preprint databases, we identified 12 sources that are cited in our dataset.

Preprint servers can be multidisciplinary, thematic, linguistics, linked to a publisher, etc. We account for this diversity in Table 3 that lists the preprint databases and the number of citations found in the dataset for each database. A total of 8460 preprint citations were extracted and used to produce the following results.

Table 3. Preprint databases



Preprint database	Discipline(s)	Created	Number of citations
arXiv	Multidisciplinary	1991	4,759
SSRN		1994	575
CogPrints		1997–2017	9
Nature Precedings		2007-2012	114
ViXra		2009	22
AAS Open Research		2018	2
preprint*			254
Cryptology ePrint Archive	Cryptography	1996	11
RePEc	Economics	1997	198
PeerJ preprint	Biology, medicine	2013-2019	137
bioRxiv	Biology	2013	2,363
PsyArXiv	Psychology	2016	15
SocArXiv	Social science	2016	1
		Total	8,460

* ‘preprint’ keyword used in the bibliography item without mentioning a specific database

[1] https://asapbio.org/preprint-servers, https://zenodo.org/record/4321522, https://en.wikipedia.org/wiki/List_of_academic_preprint_repositories, https://libguides.lib.hku.hk/preprint

Result

Progression of Citations over the Past 10 years

We have plotted the number of preprint citations that were received by the different preprint databases by year. Figure 1 shows that the overall number of preprint citations grows very rapidly since 2011 and this growth is consistent with the creation and development of new preprint repositories. For example, the bioRxiv database was created in 2013 and its presence in the citations is visible from 2014 and has been growing ever since.

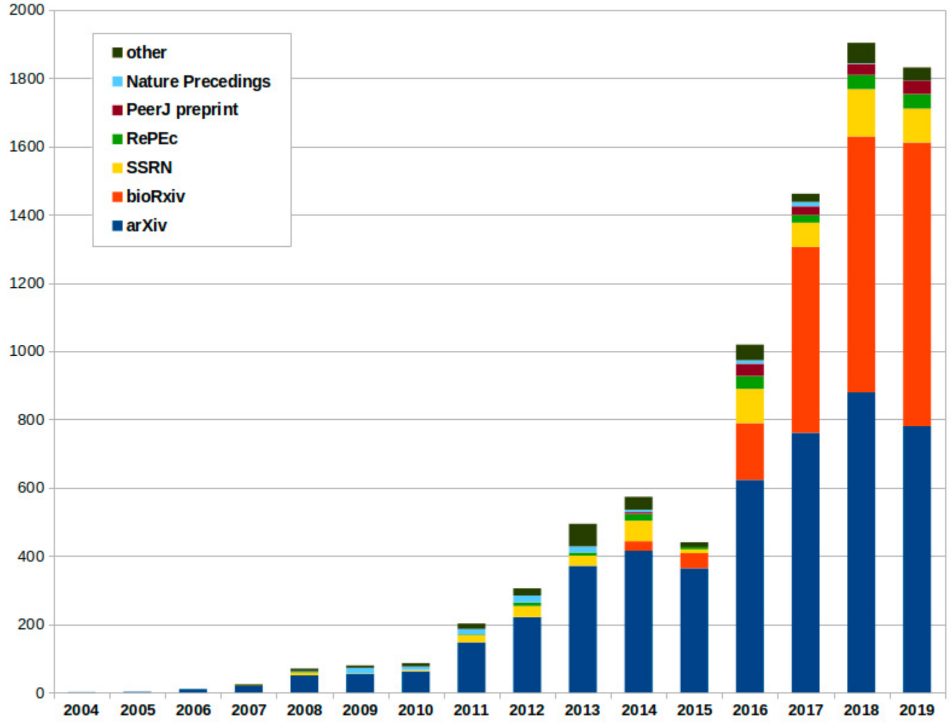


Figure 1. Progression of the number of preprint citations by year.

Frequency of Preprint Citations in the Different Sections of Articles

We have examined the positions of preprint citations with respect to the IMRaD structure in the different PLOS journals. Figure 2 presents the relative parts of preprint citations in the different sections.

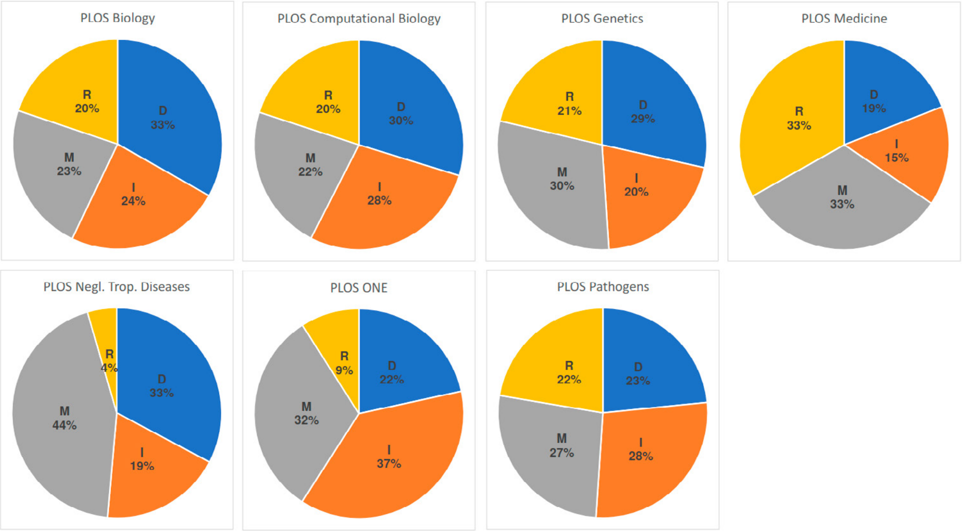


Figure 2. Preprint citations in the different sections.

One important observation is the part of preprint citations in the Method section. In fact, other studies on the distribution of citations in the IMRaD structure (Bertin et al., 2016) have shown that the Method section contains the smallest number of citations in the articles, the biggest number of citations being, naturally, in the Introduction. The fact that between 22% and 44% of preprint citations are found in the Method section is particularly interesting and means that preprint citations have distributions that are very different from the rest of the citations.

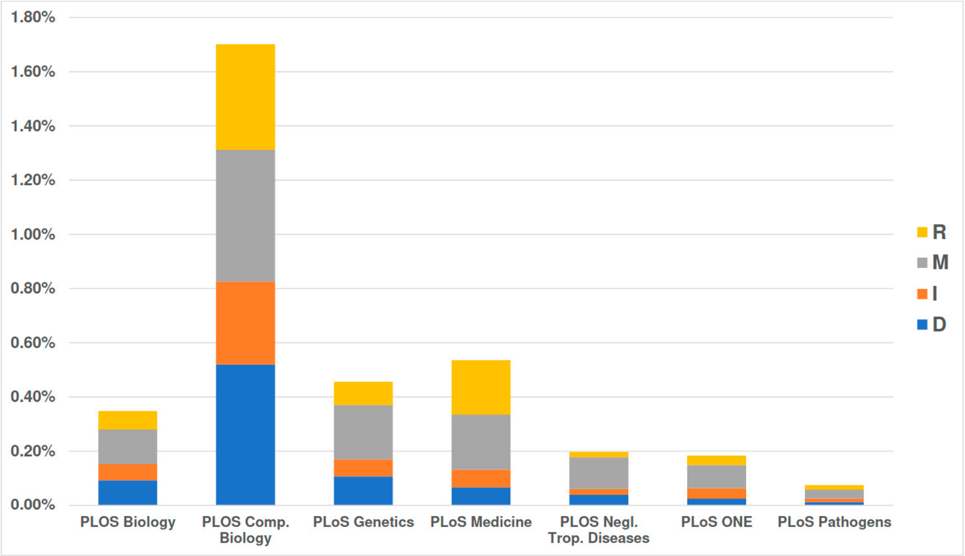


Figure 3. Preprint citations as a percentage of all citations in IMRaD.

Figure 3 shows the relative part of preprint citations as a percentage of all citations in each section and journal. Preprint citations are particularly frequent in the Method section, and somewhat less frequent in the Result and Discussion sections, except for PLOS Computational Biology. The journal PLOS Computational Biology stands out on this graph as having a very high relative number of preprint citations. Also, in this journal the parts of citations in the Method and Discussion sections are almost the same.

Distribution of Preprint Databases per Journal

The disciplinary nature of the different preprint databases is one of the factors that result in different number of citations. Figure 4 shows the distribution of the major preprint databases that were observed in the PLOS journals.

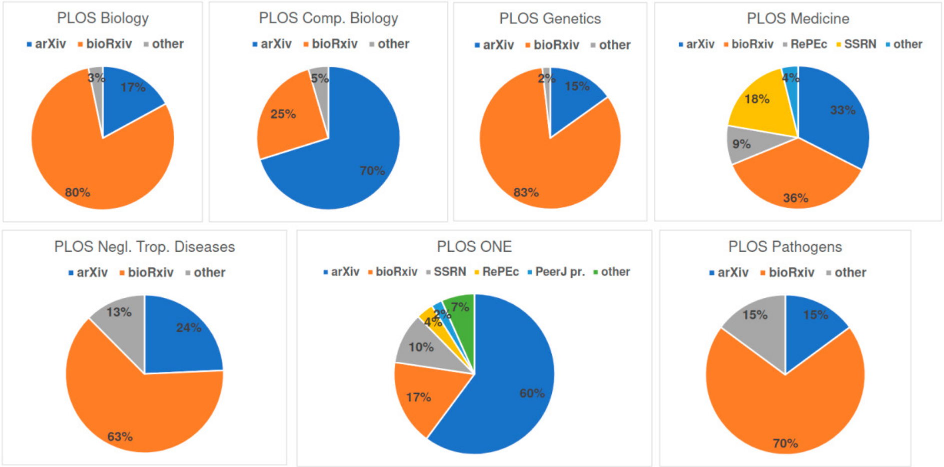


Figure 4. Distribution of preprint sources present in PLOS citations

BioRxiv is present in all PLOS journals, but it is most frequently cited in PLOS Biology, PLOS Genetics, PLOS Negl. Trop. Diseases and PLOS Pathogens. This shows clearly that the disciplinary speciality in the biomedical field leads to the wide use of this preprint database by researchers in a given field. This corroborates the results obtained by (Li,Thelwall & Kousha, 2015) who identified the point at which repositories are cited in their own domain.

Discussion and Perspectives

This paper presents an approach to study the distribution of preprint citations in scientific literature by considering their relation to the IMRaD structure of articles. We have observed the part of preprint citations in the four section types of the PLOS journals, and also the presence of the different preprint databases in the dataset.

We have shown that preprint citations gain more and more importance over the last years and this means that preprints are actively used by the community. The growth in the number of preprint citations corresponds to the creation and growth of repositories.

The position of preprint citations in the IMRaD structure shows that these citations do not follow the same distribution as the rest of the citations in the articles. In fact, an intuitive hypothesis could be that non-validated knowledge should be placed in the introduction or background of an article. Also, we could expect to find preprint citations, naturally, in the discussion section where authors could confront their results with other ongoing experiments or give new perspectives and comment on emerging approaches. However, our results show clearly that the largest part of preprint citations is to be found in the methodology section. This means that we can observe the construction of scientific knowledge that is based on research that has not been validated by the peers.

We have shown that preprint citations differ from classical citations in the sense that they do not follow the same distributions that were already established in the literature around PLOS (Bertin et al., 2016). Preprints are a new citation object to be taken into consideration and studied with the utmost attention as they represent a new paradigmatic shift. This for the following two reasons:

a) a preprint might have the quality to become article, or might simply not be intended for this purpose.
b) a preprint must be quoted with reference to the version. Indeed, a preprint can be modified over time, with no restrictions to the depth and significance of such modifications.

The next stage of this work will be to study of the contexts of citations to preprints in order to understand qualitatively the citation mechanisms behind citing preprints. Understanding citation contexts is an essential element in determining the role of preprints in terms of scientific argumentation and their contribution to the construction of new knowledge.

Acknowledgments

The research presented in this paper was supported by the French National Agency for Research grants no ANR-20-CE38-0003.

References

Abdill, R. J. and Blekhman, R. (2019). Meta-Research: Tracking the popularity and outcomes of all bioRxiv preprints, *Elife* 8 : e45133.

Berg, J. (2017). Preprint ecosystems, *Science* 357 : 1331-1331.

Berg, J. M.; Bhalla, N.; Bourne, P. E.; Chalfie, M.; Drubin, D. G.; Fraser, J. S.; Greider, C. W.; Hendricks, M.; Jones, C.; Kiley, R. and others (2016). Preprints for the life sciences, *Science* 352 : 899-901.

Bertin, M.; Atanassova, I.; Gingras, Y. and Larivière, V. (2016). The invariant distribution of references in scientific articles, *Journal of the Association for Information Science and Technology* 67 : 164-177.

Desjardins-Proulx, P.; White, E. P.; Adamson, J. J.; Ram, K.; Poisot, T. and Gravel, D. (2013). The case for open preprints in biology, *PLoS Biol* 11 : e1001563.

Fraser, N.; Momeni, F.; Mayr, P. and Peters, I. (2020). The relationship between bioRxiv preprints, citations and altmetrics, *Quantitative Science Studies* 1 : 618-638.

Fry, N. K.; Marshall, H. and Mellins-Cohen, T. (2019). In praise of preprints, *Access Microbiology* 1.

Fu, D. Y. and Hughey, J. J. (2019). Meta-Research: Releasing a preprint is associated with more attention and citations for the peer-reviewed article, *Elife* 8 : e52646.

Giles, J. (2003). Preprint server seeks way to halt plagiarists, *Nature* 426.

Hoy, M. B. (2020). Rise of the Rxivs: How Preprint Servers are Changing the Publishing Process, *Medical Reference Services Quarterly* 39 : 84-89.

Johansson, M. A.; Reich, N. G.; Meyers, L. A. and Lipsitch, M. (2018). Preprints: An underutilized mechanism to accelerate outbreak science, *PLOS Medicine* 15 : 1-5.

Kaiser, J. (2017). The preprint dilemma, *Science* 357 : 1344-1349.

Larivière, V.; Sugimoto, C. R.; Macaluso, B.; Milojević, S.; Cronin, B. and Thelwall, M. (2014). arXiv E-prints and the journal of record: An analysis of roles and relationships, *Journal of the Association for Information Science and Technology* 65 : 1157-1169.

Li, X.; Thelwall, M. and Kousha, K. (2015). The role of arXiv, RePEc, SSRN and PMC in formal scholarly communication, *Aslib journal of information management*.

Majumder, M. S. and Mandl, K. D. (2020). Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility, *The Lancet Global Health* 8 : e627-e630.

Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section, *Journal of the American Society for Information Science and Technology* 58 : 2047-2054.

Penfold, N. C. and Polka, J. K. (2020). Technical and social issues influencing the adoption of preprints in the life sciences, *PLoS genetics* 16 : e1008565.

Pinfield, S. (2001). How do physicists use an e-print archive? Implications for institutional e-print services, *D-Lib Magazine* 7 : 12.

Saier, T. and Färber, M. (2019). Bibliometric-Enhanced arXiv: A Data Set for Paper-Based and Citation-Based Tasks., *Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR) co-located with the 41st European Conference on Information Retrieval (ECIR 2019)* : 14-26.

da Silva, J. A. T. (2017a). Preprints should not be cited, *Curr. Sci* 113 : 1026-1027.

da Silva, J. A. T. (2017b). Preprints: ethical hazard or academic liberation?, *Kome* 5 : 73-80.

da Silva, J. T. (2017c). The preprint wars, *AME Med J.* 2 : 74.

Augmenting context-aware citation recommendations with citation and co-authorship history

Anubrata Bhowmick[1], Ashish Singhal[2] and Shenghui Wang[3]

[1] a.bhowmick@student.utwente.nl [2] a.a.singhal@student.utwente.nl [3] shenghui.wang@utwente.nl
University of Twente, Drienerlolaan 5, 7522 NB Enschede, (The Netherlands)

Abstract

With the increasing number of research papers being published, it has become a challenge to search for the most suitable articles for accurate referencing. Many local citation recommendation systems have begun to locate the suitable candidates by using the texts accompanying the citation suffix, along with the metadata of the target documents. Previous research has shown the positive effects from the citation relationships on such recommendations, however, the influence from the co-authorship history has not been fully investigated. In this paper, we extend the model proposed by Jeong, Jang & Park (2020) by combining the context, citation history with co-authorship information into the recommendation system. We also propose to use more domain-specific embeddings to better capture the semantics in the context. Our experiments show the positive effect of co-authorship information on citation recommendations, and that our model based on the combination of domain-specifically embedded context, the citation and the co-authorship history significantly outperforms the basic context-based recommendation model.

Introduction

Citation Recommendation is defined as the task of recommending citations from a textual content. Due to the increasing number of scientific works in recent years, and the need of citing appropriate publications when writing scientific papers, citation recommendation has become one of the most important research topics. Depending on the length of the citation context that are used for recommendation, there are two categories of methods: 1) global citation recommendation where the entire text has been used to recommend citations or abstract can be used to understand the context, and 2) local or context-aware citation recommendation where a short-length window of words surrounding the citation, as shown in Figure 1, are used as context for recommending citations.

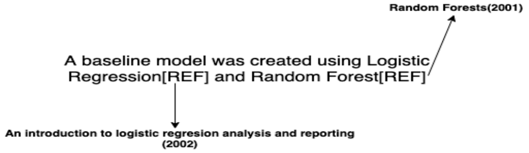

Random Forests(2001)
A baseline model was created using Logistic Regression[REF] and Random Forest[REF]
An introduction to logistic regresion analysis and reporting (2002)

Figure 1. Example of paper with citation mentions with [REF] placeholders.

There arises a vocabulary gap between the context and the corresponding cited publication which results in recommending not so high-quality publications. A publication's quality can be estimated by its previous citation relations and its authors' previous collaborations. A field breakthrough paper citing another publication speaks volumes about the cited paper quality. Similarly, a paper quality can also be determined individually by its authors, and with whom they have shared and gained knowledge by collaborating in the past. Currently existing state-of-the-art models have explored citation relations and metrics to recommend appropriate citations such as the BERT-GCN model proposed by Jeong, Jang & Park (2020). Very few have explored co-authorship relations due to the mixed views about the influence of co-authorship in this task. Ebesu & Fang (2017) employed co-author networks and found promising results and new direction for context-aware citation recommendation. In our paper, we extend the existing model of Jeong, Jang & Park (2020) with co-authorship network and domain-specific embeddings to understand if these may help in improving the performance of the existing model. By devising a way to use citation relations and co-authorship relations along with the domain-specific context embeddings, we observe a significant improvement over the existing model. Evidently, we managed to improve the Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) by 5% and Recall@k by almost 10%.

Related Work

Citation recommendation is about recommending citations given a piece of text. These systems are divided into global and local citation recommendation where global recommendation recommends the citations considering the whole publication text while local recommendation recommends citations based on local context of input text. There is a recommendation system which employs both local and global context together as described in (Muhsina & Naseer, 2019). Yang (2018) used forward and backward positions of the quotation marks of the citation in an LSTM based model to customize the context-aware citation recommendation by splitting the left and right sentences when encoding the citation context. It suggested a procedure for learning and combining LSTM cells with MLP to learn. The investigators have used the papers' author(s) and location details to encode the research model. There are other works such as Ebesu & Fang (2017) which used citation context along with authors network convolution to develop the Neural Citation Network for context-aware citation recommendation systems. This paper combined the cited and citing author network features in the citation context encoders. It becomes important to focus on co-authorship in scientific field as this dictates the quality of the research output. As per Kumar (2015), co-authorship networks generated from co-authorship meta information from various publications gives rise to a social network among the researchers which describes who is more willing to collaborate with whom influencing the publication quality. He also found that there is high correlation between co-authorships and citation behavior. It is found that co-authored papers are cited more than papers with single authors and Ding (2011) found in their study that more cited authors usually do not co-author with each other but cite each other. Lis, Yang & Ding (2009) in their study found that centrality measures from co-authorship network are highly correlated with citation counts. Biscaro & Giupponi (2014) in their study concluded that co-authorship networks' structure matters in scientific collaboration. Publications' bibliography information not only identifies the publication uniquely but also gives information about the quality of the publication. The bibliographic information such as citation count of publication, venue, year, author determines the quality of the work. These information are employed by the graph techniques which not only considers the citation count in the form edges in the graphs but also considers other bibliometric information as the nodes' features. Kipf & Welling (2016) proposed semi supervised classification with graph convolution networks to classify graph nodes by using the graph edges and nodes' features. They also introduced variational graph autoencoders to learn the latent representation of the graph.

Data Overview

For our work, we use the FullTextPeerRead dataset[1] proposed in Jeong, Jang & Park (2020). This dataset consists of peer reviews of submitted papers in top-tier venues of the ML/AI field, along with their bibliometric information. This dataset contains left and right context sentences

[1] https://bert-gcn-for-paper-citation.s3.ap-northeast-2.amazonaws.com/PeerRead/full_context_PeerRead.csv.

surrounding a citation[2] as well as the cited and the citing papers' meta information such as title, author, abstract, etc.

Table 1: Dataset description

Dataset Name	***FullTextPeerRead***[1]
Total number of papers	4,898
Total number of base papers	3,761
Total number of cited papers	2,478
Total number of citation context	17,247
Total number of unique authors	9,529
Years of papers published	2007-2017

This dataset has been split into the training and the test datasets based on the cut-off year of 2017. The training dataset consists of 3,411 papers and the test data consists of 2,559 papers. The left and right context sentences length has been trimmed to 50 words, such that sufficient information is taken from both sides of the citation.

Methodology

In order to check the effects of better context-embedding and co-authorship information, we extend the BERT-GCN model proposed in Jeong, Jang & Perk (2020) by incorporating the co-authorship network, as illustrated in Figure 2. The model consists of three components that gives embeddings for context sentences, co-authorship network and citation network. Here, the BERT transformer is used to learn the feature representation of context sentences surrounding the citations and individual Graph Convolution Networks (GCNs) are used to learn the network representations of citation network and co-authorship network. The code can be found here.[3]

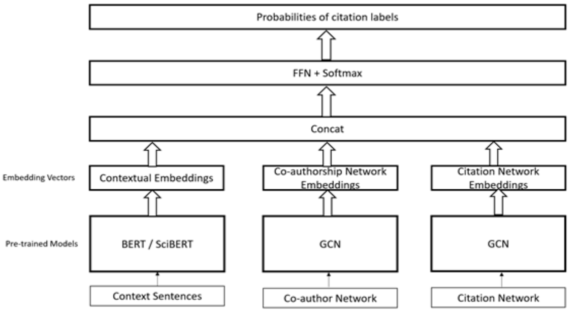


Figure 2: The BERT-GCN[2] model architecture

Context embedding

The context of a citation consists of the left and right text surrounding the cited references. We use the pre-trained BERT (Devlin, et. al., 2019) and SciBERT (Beltagy, et. al., 2019) to generate the embedding of such context. BERT in pretrained on Wikipedia articles and is capable to understand the meaning of its input due to its ability to read in both directions at once. SciBERT, an extension to original BERT is pretrained on multi-domain scientific publications and has the vocabulary which helps better understand scientific texts. As our main goal is to find suitable reference papers in computer science domain, SciBERT could be highly relevant in processing the input data from the dataset.

[2] This dataset does not consider multiple citation data where multiple papers are cited in a single sentence.
[3] https://github.com/anubratabhowmick/tf-BERT-GCN

Citation and co-authorship embedding

We generate the co-authorship network and citation network from the dataset. Two separate Graph Convolutional Network (Zhang, Hanghang & Jiejun, 2019) models are trained on these two different networks, generating two sets of embeddings for each author and each citation respectively. More specifically, for each network, we train a Variational Auto-encoder Graph Convolutional Network (VGAE) as proposed by Kipf & Welling (2017). Variational Graph Autoencoders works similarly as variational neural network auto encoders that generates stochastic based representation embeddings. We use VGAE GCN to generate the co-authorship and citation network embeddings.

Concatenation and training

The context around citations is the input to the BERT/SciBERT which on training learns these context feature representations and generates the contextual embedding. For each context sentence passed to BERT/SciBERT, the corresponding paper's citation embedding and the embedding of the last author, who is often the supervisor in the computer science domain, is concatenated with the BERT/SciBERT contextual embeddings output and passed to the feedforward network (FFN) layer. The output from this layer is then passed to the softmax activation function which then generates the probabilities of each publication being the citation for the context sentence passed as input. The BERT/SciBERT along with the feedforward layer is trained on the cross-entropy loss calculated by trying to predict the actual cited publication given the surrounding context, the corresponding citation and author embeddings as the input to the model.

Experiment Setting

The experiment has been done in eight variations, with the first four variants being trained with BERT and the last four with SciBERT, to understand how much better context embedding helps in the performance, and how much information does the respective citation and co-authorship network add to the models, in respect to the existing Jeong, Jang & Park, 2020 model. In the initial four variants, we have made a comparison of the BERT model exclusively and then, by adding the networks separately to BERT, and ending with the combined citation and co-authorship embedding added to BERT together. We repeated the same process in the next four variations by replacing the BERT with the SciBERT model.

BERT/SciBERT and GCN Setting

The context sentences are passed to BertTokenizer/SciBertTokenizer to tokenize the context sentences, respectively. The initial experiments, which only includes BERT and SciBERT separately have been trained by using these tokenized context sentences. BERT (and SciBERT) generates the contextual embeddings of the input at the [CLS] which is passed to the feedforward layer, followed by the SoftMax activation, giving the final output. The feedforward layer's input size is 768 which is equal to the output size of BERT's [CLS] token. The feedforward layer, followed by the SoftMax activation, generates the citation probabilities output of the size of 489 that is the total number of citation candidates, i.e. those that have been cited for at least 5 times in the training set. The BERT/SciBERT is being trained for 30 epochs with batch size set to 16. For training purposes, Adam optimizer (Kingma, & Ba, 2015) is used with the default parameters and the learning rate set to 2e-5.

For other settings where co-authorship networks' and citation networks' embeddings are being used, GCN has been trained for 200 epochs. The first hidden dimension in GCN for citation network training is 4837 which is total number of documents in the dataset and for co-authorship network it is 9529. The second hidden dimension in GCN is the same for both networks which is 768. The batch size for training for both networks is the total number of documents and the total number of authors in the dataset (full-batch gradient descent). For training, Adam optimizer is used with the learning rate of 0.01.

Results

We use similar metrics including Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Recall at k as mentioned in (Jeong, Jang & Park, 2020). MAP measures the average precision which reflects upon the rank position regarding the retrieval list. MRR reflects on the position of the actual result in the recommendation list. A higher MRR score indicates the higher ranks of the target citations in the recommendation list. Recall at k (R@k), where k is in [5,10,30,50,80], is used to measure the actual hit ratio in the top k recommendation results.

Table 2: MRR, MAP and R@k scores of experiments

Model Name	*MRR*	*MAP*	*R@5*	*R@10*	*R@30*	*R@50*	*R@80*
BERT Base*	0.415	0.415	0.480	0.520	0.593	0.637	0.689
BERT-GCN*(Citation)	0.418	0.418	0.486	0.529	0.604	0.649	0.699
BERT Base	0.432	0.432	0.504	0.548	0.621	0.668	0.717
BERT-GCN (Citation)	0.412	0.412	0.481	0.523	0.606	0.649	0.701
BERT-GCN (Co-Authorship)	0.439	0.439	0.505	0.556	0.632	0.677	0.728
BERT-GCN^2	0.443	0.443	0.516	0.561	0.643	0.689	0.735
SciBERT Base	0.467	0.467	0.542	0.593	0.675	0.722	0.766
SciBERT GCN (Citation)	0.467	0.467	0.547	0.592	0.677	0.722	0.772
SciBERT GCN (Co-authorship)	0.466	0.466	0.545	0.594	0.680	0.719	0.773
SciBERT-GCN^2	**0.468**	**0.468**	**0.548**	**0.598**	**0.685**	**0.733**	**0.781**

Table 2 shows our results and those (marked with *) reported by Jeong, Jang & Park (2020). Our self-implemented BERT base model outperforms the previously reported results, while the combined BERT and GCN with the citation network has slightly poor results than previously reported. This might be due to our parameter settings are not optimal. However, we can easily see the BERT-GCN with the co-authorship network performs better than both BERT-base and BERT-GCN with the citation network. A combination of both networks further improves the performance. Replacing BERT with SciBERT improves the performance significantly, even without the citation and the co-authorship network. The citation and co-authorship networks, when added to SciBERT show negligible increase in performance, but when both networks combined, the performance is again improved substantially. This suggests that co-authorship information has positive influence on context-aware citation recommendation. A combination of both the citation and co-authorship network to a base BERT/SciBERT model can certainly improve the recommendation performance.

Conclusion

Our proposed addition of the co-authorship network to the existing citation network, delivers a significant improvement in MAP, MRR and Recall@k over the existing model. This clearly shows that the co-authorship network manages to add meaningful information to the base model, and a combination with the citation network results in better citation recommendations. We can conclude that, in addition to the co-authorship network, better context embedding plays a significant role when it comes to model performance, and using proper context embeddings, along with the appropriate networks can go a long way in effective citation recommendation

There are also room for improvements in the model proposed in this paper. We have selected the last author's network embedding to be added to our model, but another viable way of using the co-author embeddings is separating the authors, increasing the dataset, and using all their embeddings separately. Although this would be computationally more expensive, there is a chance of improvement in performance over our proposed model. Another improvement could be using node features in the Graph Convolution Network, as encoding both the graph structure and the node features might improve the performance by a significant margin, while being computationally efficient, and could be worthwhile to investigate further.

References

Beltagy, I., Cohan, A. & Lo, K. (2019). SciBERT: Pretrained Contextualized Embeddings for Scientific Text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), 3615-3620. http://dx.doi.org/10.18653/v1/D19-1371

Biscaro, C. & Giupponi, C. (2014). Co-authorship and bibliographic coupling network effects on citations. *PLoS ONE*, 9(6), e99502. https://doi.org/10.1371/journal.pone.0099502

Ding, Y. (2011). Scientific collaboration and endorsement: network analysis of Coauthorship and citation networks. *Journal of Informetrics*, *5*(1), 187-203. https://doi.org/10.1016/j.joi.2010.10.008.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of The North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* volume 1, 4171-4186.

Ebesu, T. & Fang, Y. (2017). Neural Citation Network for Context-Aware Citation Recommendation. *Proceedings of the 40th International ACM SIGIR Conference of Research and Development in Information Retrieval (SIGIR '17).* ACM, New York, NY, USA, 1093-1096.

Jeong, C., Jang, S., Park, E. & Choi, S. (2020). A Context-aware Citation Recommendation Model with BERT and Graph Convolutional Networks. *Scientometrics* 124, 1907–1922. https://doi.org/10.1007/s11192-020-03561-y

Kipf, T. N. & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *Neural Processing Letters*, 1 -12.

Kipf, T., & Welling, M. (2016). Variational Graph Auto-Encoders. NIPS Workshop on Bayesian Deep Learning. arXiv:1611.07308

Kingma, D.P. & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Kumar, S. (2015). Co-authorship networks: A review of the literature. *Aslib Journal of Information Management.* 67(1), 55-73. https://doi.org/10.1108/AJIM-09-2014-0116

Ma, Y., Hao, J., Yang, Y., Li, H., Jin, J. & Chen, G. (2019). Spectral-based Graph Convolutional Network for Directed Graphs. ArXiv, abs/1907.08990.

Muhsina, V. P. & Naseer, C. (2019). A Survey on Citation Recommendation System. *International Journal of Advanced Research in Computer and Communication Engineering, 8*(1), 85-91.

Yan, E.J. & Ding, Y. (2009). Applying Centrality Measures to Impact Analysis: A Coauthorship Network Analysis. *Journal of the American Society for Information Science and Technology*, 60(10), 2017-2118. https://doi.org/10.1002/asi.21128

Yang, L., Zheng, Y., Cai, X., Dai, H., Mu, D., Guo, L. & Dai, T. (2018). A LSTM based model for personalized context-aware citation recommendation. *IEEE Access,* 6, 59618-59627. https://doi.org/10.1109/ACCESS.2018.2872730

Zhang, S., Tong, H., Xu, J. & Maciejewski, R. (2019). Graph Convolutional Networks: a Comprehensive Review. *Computational Social Network, 6, 11* https://doi.org/10.1186/s40649-019-0069-y

A systematic method for identifying references to academic research in grey literature

Matthew S. Bickley[1], Kayvan Kousha[2] and Michael Thelwall[3]

[1] *M.Bickley@wlv.ac.uk*

[2] *K.Kousha@wlv.ac.uk*

[3] *M.Thelwall@wlv.ac.uk*

[1,2,3] Statistical Cybermetrics Research Group (SCRG), University of Wolverhampton, Wulfruna Street, Wolverhampton, WV1 1LY (United Kingdom)

Abstract

Grey literature is research that has not been written with the intent to publish in a traditional journal or book. From this, and due to its unstandardised nature, its impact in academia can be difficult to identify. Research impact can be assessed in multiple ways, with citation analysis a usual method. Impact can include the citing of an output, but in some situations, cited references may be useful in assessing *'academic impact'*. Cited references in grey literature, however, may reflect *'non-academic impact'* of research such as in policy making, clinical practice or legislation. This study introduces and tests a semi-automatic method to measure cited references in grey literature with unknown standardisation of references. Metadata (lead author surname, title, year) of 2.45 million Russell Group university outputs were collected, added to known citation metadata from a 100-document sample of UK government grey literature, and then searched within each document, assessing the accuracy of 21 proposed variations of matching terms. A 'best method' is proposed (lead author surname and title in either order, maximum of 200 characters apart) to show cited references present, enabling the ability to analyse impact differences impact across subject areas and years within grey literature in future studies.

Introduction

Grey literature is a term which describes text-based documents not published in a standard academic format (e.g., books or journal articles, IGLWG, 1995). These documents are usually produced by organisations that do not focus on publishing (Schöpfel, 2010, p.11). Grey literature includes, but is not limited to, unpublished research, governmental reports, policies, conference proceedings, theses and dissertations (GreyNet International, 2019; UNE, 2019). The use of standardised reference lists in grey literature has not been widespread (Benzies et al., 2006), complicating their automated extraction and assessment.

Academic research impact has been mainly based upon counting citations from formal academic publications (e.g., journal articles) with traditional citation indexes such as Scopus or Web of Science. Nevertheless, other types of non-standard publications may be needed for monitoring the wider benefits of academic research. For instance, in the context of the UK Research Excellence Framework, impacts on "economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia" must be demonstrated (REF, 2019, p.68). Cited references in grey literature can form part of this evidence by showing specific non-academic uses of research. For instance, World Health Organization (2020) guidance on water, sanitation, hygiene and waste management for COVID-19 has cited several scientific journal articles for healthcare policy making. The historical roots of research can also be found by analysing cited references (Marx & Bornmann, 2016), which can potentially show differences in impact across various literature.

Although Altmetric (n.d.) counts references to academic research from grey literature and some previous studies have analysed documents citing grey literature in narrow contexts dating back varying years (Alberani, Pietrangeli & Mazza, 1990; Pelzer & Wiese, 2003; Cordes, 2004; Woods, Phillips & Dudash, 2020), no practical method has developed to systematically identify academic research citations in grey literature. In response, this article demonstrates a method to semi-automatically identify citations mentioned in grey literature. This would help researchers or research groups show the contribution of grey literature output within the scientific community, by showing how grey literature is both cited by other research (Bickley, Kousha & Thelwall, 2019, pp.1801-1812; Bickley, Kousha & Thelwall, 2020) and the types of studies cited. The latter of these is potentially one of the "additional possibilities for [cited reference analysis] application in scientometrics," (Marx & Bornmann, 2016) and is the focus of this study.

Research questions

The goal of this study is to assess if a predictive method can be established that reliably and robustly captures citation mentions within documents, where reference sections are not indexed or known (grey literature). UK government publications are the focus, because of the large number of freely available online documents and the authors are familiar with the locale. The research questions are:

1. Can references in non-standard grey literature be automatically identified accurately?
2. Can a comparison of newly established methods be compared to determine a best approach within the context of grey literature repository impact assessment?

Methodology

The new data extraction method uses *Publish or Perish* and *Webometric Analyst* software to gather and analyse data. It was tested on a sample of grey literature. Twenty-one heuristics to extract citations were compared using Bland-Altman analysis and plots (Bland & Altman, 1986).

To imitate a dataset held by an institution or company wishing to assess impact of their studies in a wider context, a grey literature database was needed to extract documents. The UK government website was chosen because it contains high value and varied documents that provide useful impact evidence when citing academic research. It has previously been studied to investigate the impact of these documents (Bickley, Kousha & Thelwall, 2019, pp.1801-1812; Bickley, Kousha & Thelwall, 2020), but not the references in them.

Step 1: Google Scholar search (via Publish or Perish)

Publish or Perish (Harzing, 2010) was used to identify digitised UK government reports indexed by Google Scholar during 2010-2018. Other programs such as Dimensions (Hook, Porter & Herzog, 2018) could also have been used. To generate effective searches, the '*site:*' command was used together with the '*filetype:*' command to limit the results to UK Government website documents in PDF format:

site:gov.uk filetype:pdf

The query was also amended and submitted to search for Microsoft Word documents (.doc and .docx), which are also present in the repository. Publish or Perish allows searching by year which was used to allow for discrepancies between years to be isolated so subject area differences can be analysed in a future study. From this, each query term above was searched 9 times, one for each of 2010-2018, leading to a total of 27 different search queries (9 years, 3 filetypes). This method found the URLs of about 65% (4,280 of 6,591) search results indexed in Google Scholar (4,206 PDF files, 74 Word documents; Table 1).

Step 2: Downloading UK government reports

Webometric Analyst (*Services -> Download binary or text URLs*) was used to download each document found by Publish or Perish. About 80% of the documents were successfully downloaded by Webometric Analyst (3,435 of 4,280; Table 1). The documents missed were likely due to some unstandardised PDF settings or document protection causing the automatic download of the file to fail.

Step 3: Converting PDF files to text files

The Xpdf command line tool (Glyph & Cog, 2019) through Webometric Analyst (*Text -> Convert PDF files in folder to text*) was used to automatically convert the PDF and Word documents to text, of which over 98% (3,374 of 3,425; Table 1) were successfully converted. Manual conversion of a PDF or Word document is possible, but the methods presented here are intended to form a basis for potentially larger scale studies where manual conversion is infeasible.

Table 1. Documents found using Google Scholar, extracted using Publish or Perish, downloaded using Webometric Analyst and converted using PowerShell, with totals and success ratios for each step compared to the previous step (number of documents found using Publish or Perish out of number of documents shown in Google Scholar, for example), split by year/filetype.

Year	*Filetype*	*Google Scholar*	*Publish or Perish*	*Webometric Analyst*	*Converted*
2010	pdf	801	281	242	236
	doc	11	11	10	10
	docx	1	1	1	1
2011	pdf	1070	500	406	396
	doc	11	11	11	11
	docx	0	0	0	0
2012	pdf	1030	565	462	451
	doc	14	14	11	11
	docx	2	2	1	1
2013	pdf	905	586	469	464
	doc	12	12	7	7
	docx	1	1	0	0
2014	pdf	911	612	486	479
	doc	6	6	0	0
	docx	2	2	0	0
2015	pdf	633	536	435	432
	doc	5	5	1	1
	docx	0	0	0	0
2016	pdf	496	461	361	358
	doc	1	1	1	1
	docx	4	4	2	2
2017	pdf	372	370	301	297
	doc	1	1	1	1
	docx	1	1	0	0
2018	pdf	299	295	216	214
	doc	1	1	1	1
	docx	1	1	0	0
Total		6591	4280	3425	3374
Success ratio		N/A	64.9%	80.0%	98.5%

Step 4: Identification and extraction of references

The above steps produced 3,374 plain text UK government grey literature documents. Two random samples of 50 documents were selected for initial testing. One sample had reference lists and the other did not, produced using manual checking.

The metadata and references for the 50 documents with references were manually extracted from the original PDF/Word documents to bypass automatic document processing errors (e.g., incorrect line breaks). The following metadata was also manually extracted: lead author surname, title and publication year. Reference titles with fewer than 5 words were removed (short titles may contribute to inaccurate results as the title is inherently more generic), and those with more than 10 words in the title were included but shortened to the first 10 words only to avoid matching problems (e.g., line wrapping). Table 2 shows the number of references in each document used, which only includes those with titles of 5 or more words (i.e., each document may contain more references that are not be included in the matching process).

The metadata extracted as above from the documents formed a set of potential references that a computer program could try to match with the same (or another) set of documents. This would test the ability of the program to recognise the pre-selected reference list. This would be an unrealistically small set of references to match, however. Scopus references were therefore used to expand the set, as follows. Scopus Advanced Searches were used to download the metadata (authors, title and year) for all UK Russell Group university outputs, giving 2.45 million records.

Step 5: Matching search terms using Webometric Analyst

Webometric Analyst (*Text -> Find two/three strings close together in a set of files*) was used to match the references in the expanded list with the original documents. A bespoke routine was added to allow a batch of text files to be imported and then a matching file be added to search each document in the batch for matching terms. Two versions were created (one to match 2 terms, another to match 3), and each version allowed for adjustable options; to allow for the pair (or triple) of terms to appear in any order or with a specific one first, and for the maximum distance in characters between the terms found to be classed as a match.

The different methods changed which indicators to include. Four combinations of matching terms were chosen: 1) lead author surname, title and year, 2) lead author surname and title, 3) lead author surname and year, and 4) title and year. Because many reference formats start with author names (Pears & Shields, 2019), each option was repeated with the author forced to appear before the other terms (when included). Options of 50-, 100- and 200-character maximum distances between each term were chosen by inspection of references seen in many documents. A total of 21 different methods were created using combinations of these options and used in the results, and each is numbered in Tables 2 and 3 for identification purposes.

Results

For 18 of the 21 methods (1-12 and 19-21), 49 of the 50 documents without references are always correctly predicted to have no references (Table 2, combined into one row to avoid repetition). The remaining methods (13-18) are relatively large overpredictions for these 49 documents, so although cited reference counts for each document is different, the fact that they are much larger than zero when zero is predicted by methods 1-12 and 19-21, means the exact value is unimportant. The remaining document without references is shown on its own row (Table 2) for clarity as all methods predict at least 1 cited reference when the true value is zero. For the 50 documents with known reference counts of at least 1, each document is shown on a separate row (Table 2) to illustrate the difference between the number of cited references known to exist in the document and each of the 21 methods' predicted measure.

Once calculated, Bland-Altman analysis (Bland & Altman, 1986) was used to compare the methods to the known number of manually counted references. Bland-Altman analysis is used traditionally in clinical contexts to compare two methods for estimation across multiple observations and is a simple way to estimate an agreement interval (limits of agreement), within which 95% of the differences of one method compared to the other, fall (Giavarina, 2015). However, it should be noted that these limits of agreement should be defined in advance for what the 95% central limit should be in terms of the variable being measured – it could be that the method proposed is still unacceptable if the limits are too large.

For this study, as 21 varying methods are being proposed via different combinations of indicator-pairs/triples and distances, Bland-Altman analysis can be used to show which of the 21 methods is best by finding the smallest (absolute) bias along with acceptably small limits of agreement compared to the known reference count in each document. For the best method, it is also necessary to state if the limits of agreement are acceptably small. This is because it is possible that the bias is for a method is close to zero but if the limits of agreement are too large, the method would be deemed unreliable.

In general, if a measurement is to be taken of a specific variable, but this measurement is difficult or costly, it may be of interest to know the level of another variable which can be taken easily or relatively inexpensively. Bland-Altman analysis is appropriate here as large-scale manual counting of cited references in grey literature would be time consuming if not expensive. A paired Student's t-test for the total number of references found by each method would be insufficient because it only compares group means rather than differences observation-by-observation, leading to ignoring false matches in this context.

Table 3 shows the biases and standard deviation calculated using Bland-Altman analysis for each method and ranks them within both statistics (smaller absolute bias and standard deviation is better). The most accurate method in terms of the smallest absolute bias (-0.21) and standard deviation (1.76) uses the lead author surname and title (author not necessarily first), with a maximum of 200 characters between the two (method 12). The limits of agreement are calculated similarly to a confidence interval, meaning that 95% of the predictions would be within 3.4496 cited references (3.4496=1.96*1.76, 3 or 4 rounded to the nearest integer number of references) of the true value. It is deemed that this is acceptable for a 'best-method' proposed, but further research into the distance parameter may yield improvements, reducing this uncertainty.

The same indicators used in the best method above, but with the author surname required to be first (method 9) was the second-best method for both statistics (bias=-0.28, standard deviation=1.78). All other methods have comparatively worse bias and/or standard deviation (e.g., method 21, using title and year with distance of 200, is ranked third in both statistics but the bias is more than double the best method).

Negative bias implies that the method underpredicts the number of references in each document, whereas positive bias indicates overprediction. For the 15 methods with biases and standard deviations much closer to zero (methods 1-12 and 19-21), they show underpredictions, so these methods are missing matches (false negatives) rather than including (m)any false positives. The remaining 6 methods (13-18, using author and year) are wildly overpredicting the number of references (also large standard deviations), likely due to the generic nature of a single author surname and a 4-digit integer, both of which could commonly be contained within other strings of text (e.g., "Ng 2018" is a common surname and potential year combination, but would show as a match in the string "predicting 2018 as a more fruitful year"). A space in front of the lead author surname could remove this type of false positive, but if the author surname were preceded by other characters, text markup language (such a new line delimiters), or by nothing at all (if the author surname were the first text in the document), then there would be many false negatives.

Table 2. All methods showing predicted number of references in each document against known number of references.

Combination of indicators used		*Lead Author Surname, Title and Year*						*Lead Author Surname and Title*						*Lead Author Surname and Year*						*Title and Year*		
Author surname 1st		*Yes*			*Not necessarily*			*Yes*			*Not necessarily*			*Yes*			*Not necessarily*			*N/A*		
Distance		*50*	*100*	*200*	*50*	*100*	*200*	*50*	*100*	*200*	*50*	*100*	*200*	*50*	*100*	*200*	*50*	*100*	*200*	*50*	*100*	*200*
Method number		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*
Doc. ID	*References in document*																					
1	55	51	55	55	51	55	55	51	55	55	52	56	56	1882	3032	4650	2050	3276	5030	54	54	54
2	11	7	14	17	8	15	18	7	15	18	8	16	19	2140	3083	4571	2195	3198	4757	18	19	19
3	27	5	10	19	9	18	25	19	20	22	22	23	25	1678	2523	3534	1734	2608	3658	17	23	23
4	10	6	7	8	6	7	8	6	7	8	6	7	8	459	761	1152	503	812	1243	8	8	8
5	10	10	10	10	10	10	10	10	10	10	10	10	10	521	806	1245	608	928	1430	10	10	10
6	103	75	103	103	75	103	103	76	103	103	76	103	103	2932	4415	6296	3008	4516	6508	100	100	100
7	1	1	1	1	1	1	1	1	1	1	1	1	1	81	136	255	116	181	300	1	1	1
8	3	1	3	3	1	3	3	1	3	3	1	3	3	1805	2647	3686	1906	2799	3902	3	3	3
9	10	15	15	15	15	15	16	19	19	19	19	19	19	1358	2153	3174	1496	2361	3471	15	15	16
10	3	3	3	3	3	3	3	3	3	3	3	3	3	345	579	887	387	663	1041	3	3	3
11	9	0	2	4	0	2	4	7	8	8	7	8	8	110	187	307	146	253	431	4	4	5
12	38	38	38	38	38	38	38	38	38	38	38	38	38	1210	1765	2376	1243	1820	2483	38	38	38
13	76	64	74	76	64	74	76	64	74	76	64	74	76	2270	3362	4761	2333	3468	4929	76	76	76
14	60	26	51	58	26	53	62	34	56	62	34	56	62	2867	4196	6056	2929	4322	6247	62	62	62
15	9	8	8	8	8	8	8	8	8	8	8	8	8	341	570	868	438	732	1119	8	8	8
16	61	47	57	58	47	57	58	48	58	59	48	58	59	2901	4182	5899	2971	4292	6092	58	58	58
17	6	5	6	6	5	6	6	6	6	6	6	6	6	699	927	1158	718	964	1208	6	6	6
18	29	23	28	28	23	28	28	28	30	30	28	30	30	849	1296	2063	880	1367	2195	28	28	29
19	22	22	22	22	22	22	22	22	22	22	22	22	22	1049	1578	2327	1133	1697	2510	22	22	22
20	48	38	47	49	38	47	49	38	47	49	38	47	49	1429	2184	3216	1528	2346	3437	49	49	49
21	17	3	18	18	3	18	18	3	18	18	3	18	18	425	659	1021	500	772	1166	18	18	18
22	69	55	64	69	57	64	69	58	67	69	58	67	69	2236	3176	4358	2342	3330	4558	68	69	69
23	135	66	128	132	66	128	132	67	127	131	67	127	131	2355	3343	4610	2429	3416	4711	132	132	132

24	69	46	65	68	46	65	68	47	65	69	47	65	69	1998	2882	4025	2095	3024	4236	67	68	68
25	9	5	5	5	5	5	5	5	5	5	5	5	5	496	782	1152	552	865	1275	5	5	5
26	39	35	37	37	35	37	37	35	37	37	35	37	37	960	1453	1919	1034	1538	2061	37	37	37
27	102	4	35	99	11	40	100	92	105	105	92	105	105	2468	3623	5108	2588	3789	5367	75	99	103
28	5	5	5	5	5	5	5	5	5	5	5	5	5	219	373	557	285	488	714	5	5	5
29	1	1	1	1	1	1	1	1	1	1	1	1	1	299	430	673	342	492	772	1	1	1
30	34	30	34	34	30	34	34	30	34	34	30	34	34	861	1386	2023	968	1544	2316	34	34	34
31	15	11	13	14	11	13	14	13	15	15	13	15	15	1522	2204	3059	1614	2362	3275	12	13	14
32	4	2	4	4	2	4	4	3	4	4	3	4	4	617	912	1343	696	1031	1508	3	3	3
33	9	3	4	8	3	4	8	3	4	8	3	4	8	321	597	932	382	686	1053	10	10	10
34	15	11	13	13	11	13	13	11	13	13	11	13	13	366	582	862	472	744	1095	13	13	13
35	19	7	15	19	7	15	19	7	15	19	7	15	19	621	1007	1445	663	1079	1582	19	19	19
36	12	3	6	11	3	6	11	4	8	11	4	8	11	876	1456	2238	935	1612	2530	11	11	11
37	20	11	13	15	11	13	15	11	13	15	11	13	15	1669	2524	3584	1742	2633	3734	15	15	15
38	103	49	82	95	49	83	96	51	86	99	51	86	99	2974	4320	5964	3133	4537	6265	98	98	99
39	53	31	48	51	31	48	51	38	49	51	39	49	51	1512	2301	3318	1623	2439	3563	49	49	49
40	12	8	9	9	11	12	12	10	10	10	12	12	12	467	627	924	530	733	1085	10	10	10
41	7	5	6	6	5	6	6	5	6	6	5	6	6	583	923	1261	633	993	1382	6	6	6
42	6	4	5	5	4	5	5	4	5	5	4	5	5	368	535	768	426	625	878	5	5	5
43	17	13	14	14	13	14	14	13	14	14	13	14	14	902	1554	2321	995	1684	2515	14	14	14
44	12	12	12	12	12	12	12	12	12	12	12	12	12	625	922	1337	687	1034	1455	12	12	12
45	83	65	79	79	65	79	79	65	79	79	65	79	79	2467	3738	5417	2559	3880	5672	79	79	79
46	84	0	1	42	0	7	46	20	80	80	20	80	80	2669	4091	5569	2764	4256	5757	31	69	80
47	15	12	14	15	12	14	15	13	15	15	13	15	15	939	1325	1907	990	1406	2029	15	15	15
48	107	0	1	66	0	5	68	38	105	105	38	105	105	2559	4086	5907	2642	4255	6130	42	86	102
49	22	11	19	22	11	19	22	11	19	22	11	19	22	1545	2156	2884	1630	2281	3076	22	22	22
50	6	3	5	6	3	5	6	3	5	6	3	5	6	1157	1523	2122	1171	1545	2157	6	6	6
49x docs.	0	0	0	0	0	0	0	0	0	0	0	0	0	All >0	All >0	All >0	All >0	All >0	All >0	0	0	0
1x doc.	0	1	1	1	1	1	1	1	1	1	1	1	1	480	724	1099	571	869	1322	1	1	1
Total	1692	957	1310	1556	974	1340	1579	1165	1605	1664	1173	1612	1671	76396	114983	165368	81540	123089	177634	1496	1612	1648

Table 3. All methods showing bias (mean of known and method-reported number of references), standard deviation and rank of each (smaller absolute bias and standard deviation is better).

Method number	*Bias*	*Standard deviation*	*Bias rank*	*Standard deviation rank*
1	-7.35	19.01	15	15
2	-3.82	15.04	11	13
3	-1.36	6.04	8	8
4	-7.18	18.66	14	14
5	-3.52	14.18	10	12
6	-1.13	5.61	7	7
7	-5.27	13.27	13	11
8	-0.87	2.67	6	5
9	-0.28	1.78	2	2
10	-5.19	13.27	12	10
11	-0.80	2.64	=4	4
12	-0.21	1.76	1	1
13	747.04	809.24	16	16
14	1132.91	1209.32	18	18
15	1636.76	1720.39	20	20
16	798.48	835.59	17	17
17	1213.97	1250.06	19	19
18	1759.42	1783.49	21	21
19	-1.96	8.85	9	9
20	-0.80	3.04	=4	6
21	-0.44	1.78	3	3

Bland-Altman plots of the methods can illustrate the difference between the known and predicted number of references against mean of the two. A ‘good’ prediction method (Figure 1) shows a scatter plot which has similar spread across all parts of the x-axis (mean of known and method-reported number of references), ideally with points being close to zero on the y-axis (difference between known and method-reported number of references), meaning narrow limits of agreement. It follows that this method would then have low bias and standard deviation (limits of agreement), meaning the method presented is a good predictor of the true measure.

An unsuitable method (Figure 3) would show ‘fanning-out’ along the axis from left to right. Here, as all measures are positive integers and the worst of the prediction methods presented wildly overpredict rather than underpredict, ‘one-sided fanning-out’ is shown but leads to the same conclusion; larger bias and limits of agreement are visible.

Figures 1-3 show Bland-Altman plots from the best ranked method (12), the worst ranked of those with relatively close-to-zero bias/standard deviation (method 1) and the worst ranked overall (method 18). Bland-Altman plots for the other methods (Figures 4-21) are shown in the online Appendices (Appendices 1-18, Bickley, Kousha & Thelwall, 2021).

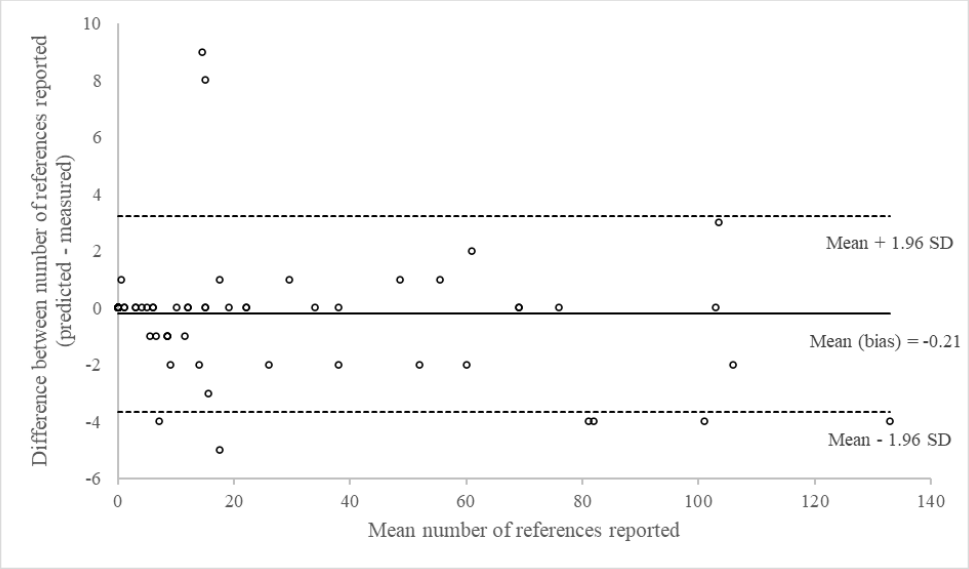


Figure 1. Bland-Altman plot of agreement between method 12 (lead author surname and title, author not necessarily first, distance 200) and known reference count of 100-document sample, the best method proposed.

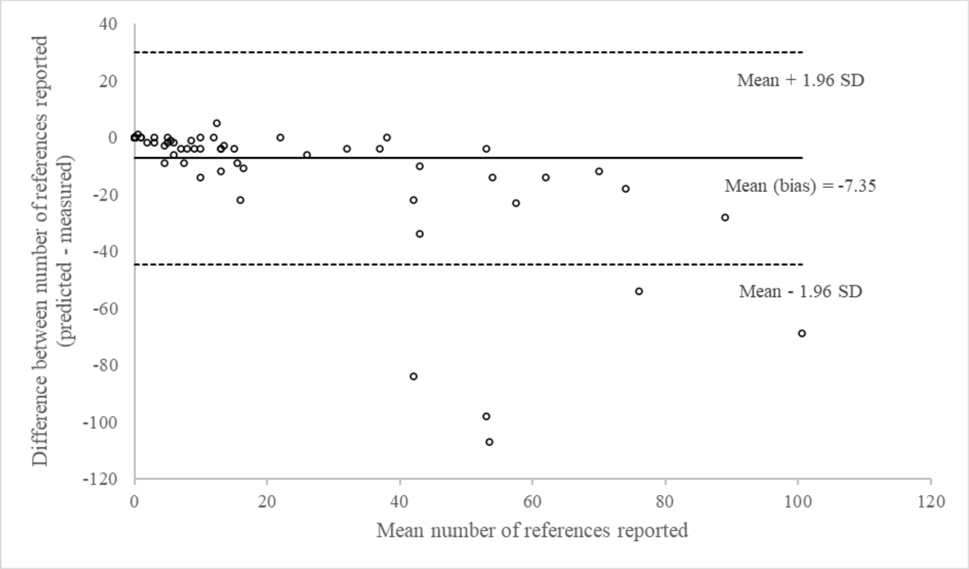


Figure 2. Bland-Altman plot of agreement between method 1 (lead author surname, title and year, author first, distance 50) and known reference count of 100-document sample, showing larger bias/limits of agreement (bias ± 1.96 SD) to Figure 1.

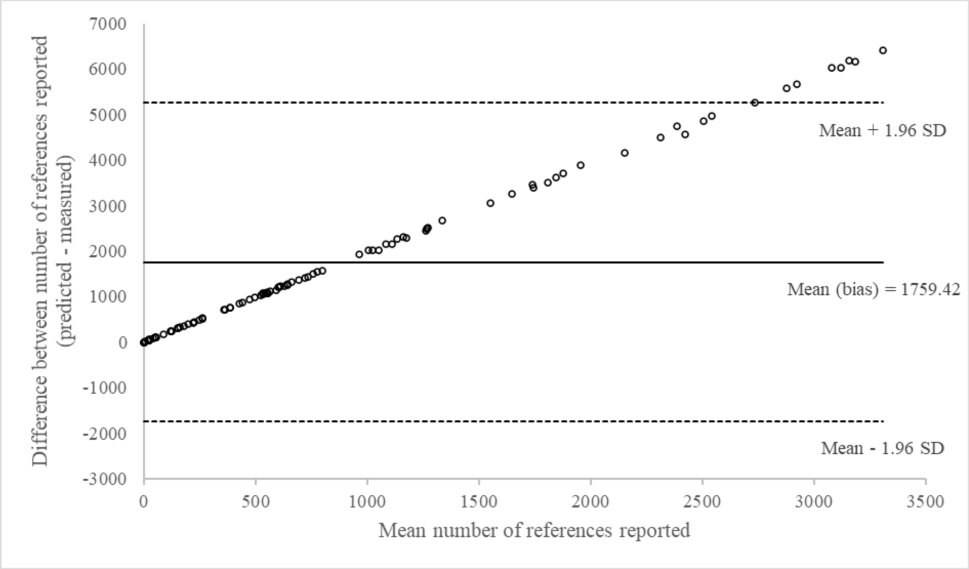


Figure 3. Bland-Altman plot of agreement between method 18 (lead author surname and year, author not necessarily first, distance 200) and known reference count of 100-document sample, demonstrating a poor method with 'one-sided fanning-out'.

Discussion and limitations

The best method found here does not necessarily represent the new 'gold-standard' to measure impact of grey literature upon standard academic output but is an accurate and reliable approach for use in further research. In this way, the combination of options above that lead to the smallest difference (bias) and variation across all records is the aim, allowing the conclusion that the best method proposed has both high recall and precision. The best method was deemed to having an acceptably low standard deviation/limits of agreement, but further research into the methodologies here may present a way at reducing this variability while keeping the bias small.

The distances between indicators chosen were arbitrary. A pattern in the result shows that of the more feasible approaches (methods 1-12 and 19-21), the larger the distance between matching terms, the smaller the absolute value of the bias and the standard deviation, hence more accurate and precise. It is plausible that a larger distance may obtain a more accurate method, although this was not tested in this research.

It is possible that due to the short nature of some lead author surnames and years, they could be contained within other strings, leading to false positives. The best method proposed here does not make use of the year of a reference but does include lead author surnames. An extended method could include looking for a preceding line break (in the case of most reference lists where each reference starts with the lead author surname) and/or a following comma (a common format in references, Pears & Shield, 2019). However, this brings extra complexity to the method and may cause missed matches if the grey literature uses a non-standard format.

The results are limited to using a single case study (UK government grey literature publications matched with Russell Group university academic output). It is possible that other grey literature repositories may have more standardised ways of reporting references within documents or recording these within metadata (e.g., Scopus), but this method has been designed to be applicable to any text documents, regardless of definition or knowledge of content.

Although this study has been limited to a sample of documents taken from a single grey literature repository, the inclusion of tools such as Publish or Perish and Webometric Analyst allow for much larger scale studies to use the methodology presented here. This process has been undertaken in a later study, with promising results assessing impact of an entire repository and allowing for subject areas differences to be seen.

Conclusions

In answer to the research questions, it is possible to count the number of references in a batch of files in a semi-automated way by using the combinations of indicators above, hence identifying reference sections. This was possible here as a large list of metadata was collected from Scopus – the hardest part of this method due to the manual collection. If these tests were to be repeated by other users, it may be useful to find an automated way of collecting data like this, and to make sure the origin of potential matches is suitable (i.e., US-based journals/books if undertaken on US-based repositories).

The exact number of citations here may be trusted as the lead author surname, title and year of the known matches were deliberately included in the matching process to check if the method is reliable for both false positives and negatives. In addition, 21 different combinations of options were tested extensively to determine a best approach. The ultimate method proposed (lead author surname and title in either order, a maximum of 200 characters apart) has low underestimation of reference count and can be treated as effective. However, when performing on a dataset with potentially any number of references, any method is likely to present many missed matches due to the unknown nature of potential citations – all academic output from all recorded years would have to be included in the list of matching terms to remove this problem. It may be more appropriate to use the method proposed here to assess impact in terms of proportion between different subject areas rather than pure numbers or if the scope of academic references used is small and known. A future project using this method on a wide scale within different subjects is underway by the authors of this paper.

Researchers wishing to assess impact of their own grey literature should endeavour to create DOIs where appropriate so persistent identifiers already used to assess impact by other mediums (Altmetric, n.d.) can be extended into the grey literature realm.

Appendices

Appendices 1-18, as mentioned in this study, can be found in the online Appendices (Bickley, Kousha & Thelwall, 2021).

References

Alberani, V., Pietrangeli, P. D. C. & Mazza, A. M. (1990). The use of grey literature in health sciences: a preliminary survey. *Bulletin of the Medical Library Association*, 78(4), 358.

Altmetric (n.d.). Altmetric for Institutions. Retrieved April 26, 2019 from: https://www.altmetric.com/audience/institutions/

Benzies, K.M., Premji, S., Hayden, K.A. & Serrett, K. (2006). State-of-the-Evidence Reviews: Advantages and Challenges of Including Grey Literature. *Worldviews on Evidence-Based Nursing*, 3: 55-61. https://doi.org/10.1111/j.1741-6787.2006.00051.x

Bickley, M.S., Kousha, K. & Thelwall, M. (2019). Can the impact of grey literature be assessed? An investigation of UK government publications cited by articles and books, in Catalano, G., Daraio, C., Gregori, M., Moed, H.F. and Ruocco, G. (eds.) *17th International Conference on Scientometrics & Informetrics, ISSI2019: Proceedings, Volume II*. Italy: International Society for Scientometrics and Informetrics/Edizione Efesto, 1801-1812.

Bickley, M.S., Kousha, K. & Thelwall, M. (2020). Can the impact of grey literature be assessed? An investigation of UK government publications cited by articles and books. *Scientometrics*, 125(2): 1425-1444. https://doi.org/10.1007/s11192-020-03628-w

Bickley, M.S., Kousha, K. & Thelwall, M. (2021). *A systematic method for identifying references to academic research in grey literature (Appendices)*. https://doi.org/10.6084/m9.figshare.12026820

Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476), 307-310. https://doi.org/10.1016/s0140-6736(86)90837-8

Cordes, R. (2004). Is grey literature ever used? Using citation analysis to measure the impact of GESAMP, an international marine scientific advisory body. *Canadian Journal of Information and Library Science*, 28(1), 49-70.

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, 25(2): 141-151. https://doi.org/10.11613/BM.2015.015

Glyph & Cog, LLC (2019). *Download Xpdf and XpdfReader*. Retrieved March 24, 2019 from: https://www.xpdfreader.com/download.html

GreyNet International (2019). *Document Types in Grey Literature*. Retrieved January 4, 2019 from: http://www.greynet.org/greysourceindex/documenttypes.html

Harzing, A.W.K. (2010). *The publish or perish book*. Melbourne: Tarma software research.

Hook, D.W., Porter, S.J. & Herzog, C. (2018). Dimensions: Building Context for Search and Evaluation. *Frontiers in Research Metrics and Analytics*. 3:23. https://doi.org/10.3389/frma.2018.00023

Interagency Gray Literature Working Group (IGLWG, 1995). *Gray Information Functional Plan (GIFP)*. Retrieved January 7, 2019 from: https://apps.dtic.mil/dtic/tr/fulltext/u2/b300928.pdf

Marx, W. & Bornmann, L. (2016). Change of perspective: bibliometrics from the point of view of cited references—a literature overview on approaches to the evaluation of cited references in bibliometrics. *Scientometrics*, 109: 1397–1415. https://doi.org/10.1007/s11192-016-2111-2

Pears, R. & Shields, G.J. (2019). *Cite them right: the essential referencing guide*. Macmillan International Higher Education.

Pelzer, N. L. & Wiese, W. H. (2003). Bibliometric study of grey literature in core veterinary medical journals. *Journal of the Medical Library Association*, 91(4), 434.

Research Excellence Framework (2019). Guidance on submissions, 68. Retrieved March 26, 2020 from: https://www.ref.ac.uk/media/1092/ref-2019_01-guidance-on-submissions.pdf

Schöpfel, J. (2010). Towards a Prague Definition of Grey Literature. *Twelfth International Conference on Grey Literature: Transparency in Grey Literature*. Prague, Czech Republic, 6-7 December 2010. Grey Tech Approaches to High Tech Issues, 11-26.

Thelwall, M. (2012). *Introduction to Webometric Analyst 2.0: A Research Tool for Social Scientists*. Retrieved February 18, 2020 from: http://lexiurl.wlv.ac.uk/searcher/IntroductionToWebometricAnalyst2.doc

University of New England (UNE, 2019). *Grey literature*. Retrieved January 4, 2019 from: https://www.une.edu.au/library/support/eskills-plus/research-skills/grey-literature

Woods, S., Phillips, K., & Dudash, A. (2020). Grey literature citations in top nursing journals: a bibliometric study. *Journal of the Medical Library Association*, 108(2), 262. https://doi.org/10.5195/jmla.2020.760

World Health Organization (WHO, 2020). *Water, sanitation, hygiene and waste management for COVID-19*. Retrieved March 25, 2020 from: http://www.dwi.gov.uk/2020-03-03%20WASH-IPC_EN.pdf

Research funding and scientific performance: measuring the impact of a PhD scholarship program

Adriana Bin[1], Sergio Salles-Filho[2], Ana Carolina Spatti[3], Jesús Pascual Mena-Chalco[4] and Fernando Antonio Basille Colugnati[5]

[1] *adribin@unicamp.br*
University of Campinas, School of Applied Science (Brazil), Pedro Zaccaria Street, 1300, Limeira - SP

[2] *sallesfi@unicamp.br*
University of Campinas, Institute of Geosciences (Brazil), Rua Carlos Gomes, 250, Campinas – SP

[3] *anaspatti@ige.unicamp.br*
University of Campinas, Institute of Geosciences (Brazil), Rua Carlos Gomes, 250, Campinas – SP

[4] *jesus.mena@ufabc.edu.br*
Federal University of ABC, Center for Mathematics, Computation and Cognition (Brazil), Av. dos Estados, 5001, Santo André - SP.

[5] *fernando.colugnati@medicina.ufjf.br*
Federal University of Juiz de Fora, School of Medicine (Brazil), Av. Eugênio do Nascimento, s/n° - Dom Bosco, Juiz de Fora - MG

Abstract

This paper presents an impact evaluation of São Paulo Research Foundation (FAPESP)'s Doctoral Scholarship Program. The objective is to assess whether the program impacts the scientific production of its beneficiaries. The study population included all the individuals who applied for PhD scholarships from FAPESP between 2003 and 2017 and were either granted or rejected. The evaluation employed a quasi-experimental design. The treatment group comprehended individuals whom FAPESP awarded and completed their PhD degree. The control groups comprehended individuals who were rejected in their FAPESP application but completed their PhD with scholarships from other funding agencies or those without funding support. Data collection was carried out on secondary bases from different sources (including Scopus and Google Scholar for citations and H-index). For the analysis, we used exact matching with the Propensity Score (PS) as matching distance. As a result, we find that the number of publications, number of citations, and H-index is higher for former FAPESP scholarship holders than for those rejected. The effect is even more significant compared to those who did not have the financial support of research agencies during their PhD studies. Therefore, findings demonstrate that scholarships for doctoral students drive the production of qualified scientific knowledge.

Introduction

The importance of scientific research for creating new knowledge, the generation of innovation, and promoting the competitiveness of regions and countries are recognized by academics and policymakers (Benavente et al., 2012; Lin et al., 2014). In this context, the existence of qualified researchers is seen as a way to expand the scientific, technological, and innovation skills of a region and to promote economic and social development (Leitch, 2006; Halse & Mowbray, 2011; Salter & Martin, 2001; Tremblay, 2005; Neumann & Tan, 2011). In Stephan's words (1996, p. 1199), "science is a source of growth."

The historical governmental interest illustrates this understanding materialized in public policies aimed at funding research (Benavente et al., 2012; Lin et al., 2014; Langfeldt; Bloch; Sivertsen, 2015) and training qualified human resources, particularly in master and PhD levels (Stephan, 1996; Ali; Bhattacharyya & Olejniczak, 2010).

There is no doubt that graduate students are considered an essential part of the academic workforce. OECD. Stat data points out about 278 thousand doctoral graduates in 2018 in OECD countries. Some of the non-OECD economies, such as Brazil, also have expressive numbers –

23 thousand doctoral graduates in the same year. Considering that doctoral studies take some years, the number of students enrolled in PhD programs worldwide is quite expressive.

Larivière (2012) reinforces the importance of this workforce with evidence of the significant contribution of PhD students to the advancement of knowledge through their participation in the publication process. The expression "on the shoulders of students" used by the author brings the real meaning of this contribution and raises a complimentary issue regarding the guarantee of adequate conditions through research funding for those students to dedicate themselves to their doctoral research.

The association of these two factors – research funding and scientific productivity and impact – has been investigated by previous studies regarding doctoral and post-doctoral programs in different countries and particular knowledge fields and time intervals. In general, findings show a positive relationship between funding and performance of those that has been supported (Goldsmith et al., 2002; Bornmann et al., 2010; Jastrob, Bukstein y Güimil, 2012; Larivière, 2013; Meyer and Bührer, 2014; Horta et al., 2018; Belavy et al., 2020).

This positive performance of alumni in scientific production is mainly explained because scholarships allow their beneficiaries to focus entirely on their doctoral research during the program. On the other hand, other studies (Böhmer & Von Ins, 2009 and Langfeldt et al., 2015) did not find the same positive relationship between funding and scientific performance.

Although a doctoral student's whole experience cannot be diminished to scientific publications, this is undoubtedly a critical output and a fundamental element in young researchers' training and attaining research excellence, as explored by Lindahl et al. (2020). Thus, further investigation about the association of funding and scientific performance must be conducted, particularly in developing countries such as Brazil, where this kind of studies are scarce and, conversely, threats regarding governmental funding to science, technology, and innovation, and particularly doctoral scholarships, are abundant (Angelo, 2017; Escobar, 2019; Andrade, 2019). This article fills this gap by providing insights on the impact of research funding on PhD holders' scientific performance in the Brazilian context.

Our purpose was to evaluate the impact of São Paulo Research Foundation's (FAPESP)'s Doctoral Scholarship Program. FAPESP is the largest state research funding agency in Brazil, accounting for almost 40% of the total financial support given annually to research activities in the state. Established in 1962, the FAPESP Scholarship Program's main objective is to develop scientific or technological research and new staff training for the research system of the state of São Paulo.

The research question that guides this article is whether the program impacts the scientific performance of its beneficiaries. A previous study on this same program (Bin et al., 2015 and Bin et al., 2016) brings forth some conclusions in convergence with the presented literature. However, it is limited for not including in the comparison individuals who have completed their PhD degree without any scholarship, using self-reported primary data and employing impact factor (IF) and not citations as a proxy of scientific impact.

The present research advances on these three fronts, expanding the control group in the evaluation and employing secondary data from the integration of different sources of information, seeking to test the hypothesis that the impact of the scientific production of the former FAPESP awardees is more significant than those who did not have a FAPESP scholarship. Secondary data also enabled analyze of circa 21 thousand individuals from all knowledge fields and compare scientific impact within two different databases (Google Scholar and Scopus).

Our paper is organized as follows. Section 2 outlines the Brazilian background regarding the research-funding system for PhD students. Section 3 outlines data and methods. Section 4 presents our findings and discussion. Finally, we bring some conclusions and implications of the study.

Background

Brazil's research funding system for PhD students is organized at both federal and state levels. At the federal level, scholarships are awarded by the Coordination for the Improvement of the Higher Education Personnel (CAPES) and the National Council for Scientific and Technological Development (CNPq). State research funding agencies award these scholarships at the state level, among which FAPESP is by far the largest.

The federal agencies CAPES and CNPq historically used a decentralized institution-based system to award scholarships. These agencies allocate quotas to universities and other research organizations based on established criteria of their institutional performance in a national ranking of graduate programs: institutions with better-ranked programs receive more scholarships than the others. The institutions themselves then follow non-standardized internal awardee- selection processes.

In its turn, FAPESP uses peer review, not the institutional system, to award scholarships based on proposals presented by applicants regularly enrolled at institutions in São Paulo State. The criteria employed in the peer review are the student's academic excellence, the supervisor's experience and expertise, and the quality of the project (objectives, theoretical base, methodology, and feasibility).

PhD students from São Paulo State can apply for a scholarship under the institutional quota or directly to FAPESP. This decision is often taken with their supervisors and considers the availability of institutional quota and the chances to be awarded. FAPESP scholarships mainly differ from CAPES/CNPq's in terms of resource amount - although, with variations in the period, the former's value is about 50% higher than that of the others. There is also a frequent follow-up of the awardees from the peers based on regular reports.

Data and Methodology

The study population was composed of all the individuals who applied for FAPESP's doctoral scholarships between 2003 and 2017 and had their requests either granted or rejected, but with the added condition that they had concluded their PhD degree.

The central database containing the data of individuals who applied for FAPESP in the evaluated period was made available by the Foundation. In addition to this base, two others were used: CAPES, which allowed identifying all students linked to Postgraduate programs in the evaluated period and who completed their PhD degree, and CV Lattes, with information from the academic curricula registered on the Lattes Platform, making it possible to identify whether or not these individuals had a scholarship from another funding agency during their doctorate.

A quasi-experimental evaluation design was employed to test our hypothesis. The treatment and the two control groups, described in the sequence, were made using exact matching with the Propensity Score (PS) as matching distance. The PS is the probability that a given individual belongs to the treatment group.

- Treatment group: individuals who have been awarded with a doctoral scholarship from FAPESP and who have completed their PhD (from now on identified as "FAPESP awardees");
- Control group 1: individuals who had their scholarships request rejected from FAPESP and who completed PhD with scholarships from other agencies – CAPES/CNPq (from now on referred to as "other awardees");
- Control group 2: individuals who had their scholarships request rejected from FAPESP and completed PhD without any funding support (from now on referred to as "no awardees").

The observable baseline control variables used in the pairing, the support domain for PS estimation, were:

- The starting year of the different academic stages of the individuals (undergraduate, master and PhD);

- The field of knowledge of doctoral research of the individuals;
- The institution and city in which the PhD was carried out (since the quality of institution might play a significant role);
- The occurrence of PhD funding support and, if positive, the type of scholarship (FAPESP or CAPES/CNPq); and
- The former scientific production of the individuals.

Two sets of PS were estimated, one for each control group, providing two different effective samples for treatment impact evaluation. Table 1 shows the size of the eligible population for each quasi-experiment, considering the variables listed above.

Table 2. The eligible population for each quasi-experiment

	FAPESP Awardees	*Other Awardees*	*No Awardees*	*Total*
PhD	12.618 (59,89%)	6.814 (32,34%)	1.635 (7,76%)	21.067 (100%)

Source: Research data.

The comparative analysis of former awardees' and no awardees' scientific production was based on data provided by individuals in their CV Lattes and organized in four categories, namely: (i) complete papers published in journals; (ii) published books; (iii) chapters of published books; and (iv) complete papers in conference proceedings.

Two types of analysis of scientific production were undertaken. The first considered the number of publications in these four categories. For this, we considered the average number of publications by individuals before and after completing their PhD. The option to use all these categories is based on the finding that different areas of knowledge have very different dynamics in terms of the primary vehicles used to communicate the results of research. Thus, the use of the category "complete articles published in journals" - most commonly used in bibliometric studies - could impose a limit to assess the behavior of areas in which articles in events, books, and book chapters are typical.

The second type of analysis was based on the impact of complete articles published in journals, measured by the number of citations and the H-index.[i]. Citation data and H-index were extracted from two different bases for comparison: Google Scholar and Scopus. Means of citations and H-index were calculated before and after the end of the doctoral studies, and the FAPESP scholarship effect indicates the scholarship's influence on these indicators.

Once the matching was done and data collected, the FAPESP scholarship effects were estimated, using the Generalized Linear Models (GLM) approach (Nelder & Wedderburn, 1972), choosing the link function and distribution family appropriate to the type of outcome/output indicator. In this article, since all indicators are based on counting, the model used was the Quasi-Poisson. This one uses Poisson distribution and log link function. The estimated effect is the ratio between the average occurrences in each group, with 1 having no effect. The term "quasi" is since the variance of the data is estimated proportionally to the average and is not exactly the average, as in the usual Poisson. This feature allows for the modeling of data that breaks this assumption of "mean equals variance," both for more (overdispersion) and for less (subdispersion). The data were analyzed using the software R.

For each bibliometric indicator, the model estimated the FAPESP (treatment) effect adjusted for the lagged indicator before the grant to control the regression to the mean bias. For instance, the number of papers after the scholarship was adjusted for the total number of articles before the scholarship.

Results

Figure 1 compares the number of papers in conference proceedings and journals published by former FAPESP scholarship awardees and individuals who received scholarships from another agency and the number of book and book chapters, in different knowledge fields and in general. The dot on the graph indicates the effect, and the lines indicate the confidence interval.

The so-called "FAPESP scholarship effect" indicates the extent to which the scholarship influenced the individuals' number of publications. For interpretation purposes, values equal to one indicate no effect. Values less than one indicate that former FAPESP awardees' performance is lower than that of their peers who had a scholarship from another agency, while values greater than one indicate that the performance of former FAPESP awardees is superior to that control group.

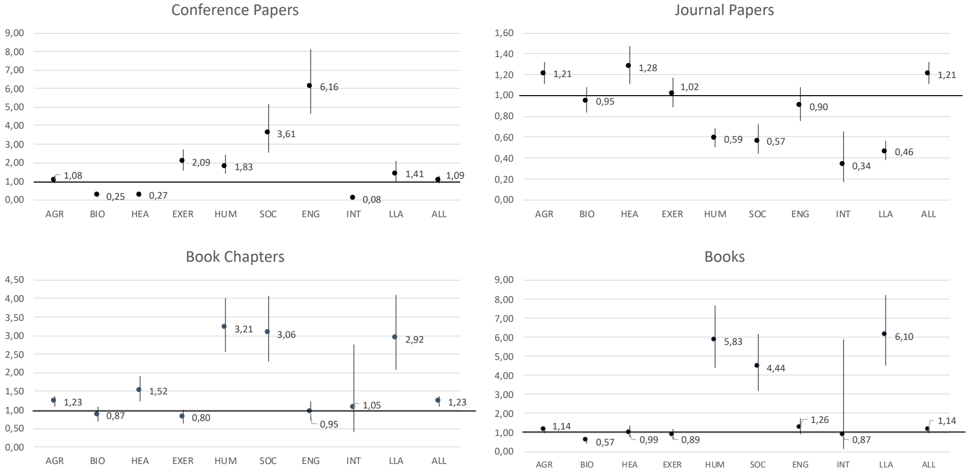


Figure 1. Publication ratio between FAPESP awardees and other awardees (number)

Legend: AGR: Agricultural Sciences; BIO: Biological Sciences; HEA: Health Sciences; EXER: Exact and Earth Sciences; HUM: Human Sciences; SOC: Social Applied Sciences; ENG: Engineering; INT: Interdisciplinary; LLA: Linguistics, Literature, and Arts; ALL: General.

Source: Research data.

It is possible to state that the former PhD FAPESP awardees publish more articles in journals, book chapters, and books than the former PhD awardees from other agencies. The effects are 21%, 23%, and 14%, respectively.

There are significant differences in knowledge fields. The FAPESP scholarship effect is positive for conference papers in all areas of knowledge except for Biological Sciences, Health Sciences, and the Interdisciplinary field. The main highlights are for Engineering and Social Applied Sciences, in which the FAPESP effect is 516% and 261%, respectively.

For papers in journals, the effects are positive for Health Sciences (28%) and Agricultural Sciences (21%) and neutral and negative for all other areas. For book chapters, the effects are very favorable for Human Sciences, Social Applied Sciences and Linguistics, Languages, and Arts and Agriculture and Health. For Biological Sciences, the effects are negative. Finally, for books, the same thing happens as for book chapters: the effects are very favorable for Human Sciences, Social Applied Sciences, Linguistics, Languages and Arts and Agriculture. For Biological Sciences, the effect is negative.

The following figure brings the same information as the others but now comparing former FAPESP scholarship holders with individuals who did not receive a doctoral scholarship.

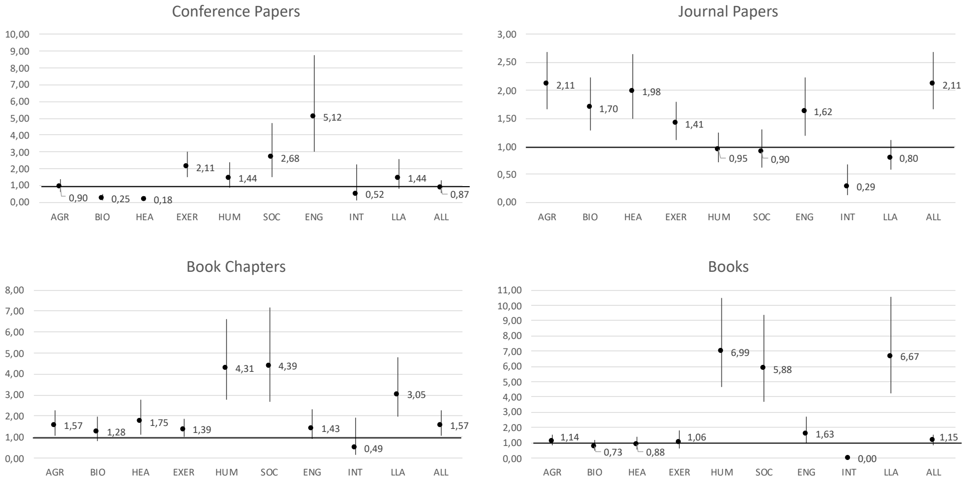


Figure 2. Publication ratio between FAPESP awardees and no awardees (number)

Legend: AGR: Agricultural Sciences; BIO: Biological Sciences; HEA: Health Sciences; EXER: Exact and Earth Sciences; HUM: Human Sciences; SOC: Social Applied Sciences; ENG: Engineering; INT: Interdisciplinary; LLA: Linguistics, Literature, and Arts; ALL: General.

Source: Research data.

Once again, the former FAPESP PhD fellows publish more articles in journals and book chapters than individuals who have done their PhD without a scholarship. These are essential effects: 111% and 57% respectively.

As expected, there are also quite different behaviors between knowledge fields. There are negative effects in only two cases: Biological Sciences and Health Sciences for conference papers and the Interdisciplinary field for papers in journals. In all other categories in which the measures are statistically significant, the effects are positive.

As can be seen, FAPESP PhD scholarship has positive effects in terms of the number of publications in vehicles of great importance for each of the knowledge fields, generally journal papers for Agricultural Sciences, Biological Sciences, Health Sciences, Exact and Earth Sciences and Engineering, and book chapters and books for Human Sciences, Social Applied Sciences and Linguistics, Literature, and Arts. Negative FAPESP PhD scholarship effect in conference papers compared to no awardees show that the latter concentrate their efforts in conferences rather than journals, which could be considered a relatively more straightforward path for disseminating research results.

The following Table (Table 2) presents impact indicators corresponding to the second type of analysis: H-index and citations.

Table 2. Citations and H-index

	Google Scholar H-Index Comparison: without scholarship			*Google Scholar H-Index Comparison: with scholarship*			*Google Scholar Citations Comparison: without scholarship*			*Google Scholar Citations Comparison: with Scholarship*		
	Ratio effects	*[confint]*	*p*	*Ratio effects*	*[confint]*	*p*	*Ratio effects*	*[confint]*	*p*	*Ratio effects*	*[confint]*	*p*
Mean on controls	0.55	[0.44, 0.69]	<0.001	1.62	[1.51, 1.73]	<0.001	11.70	[8.14, 16.82]	<0.001	52.87	[40.65, 68.77]	<0.001
FAPESP Effect	3.77	[2.99, 4.76]	<0.001	1.36	[1.27, 1.47]	<0.001	5.72	[3.94, 8.32]	<0.001	1.27	[0.97, 1.66]	0.088
Before PhD	1.21	[1.16, 1.27]	<0.001	1.15	[1.11, 1.18]	<0.001	1.00	[1.00, 1.00]	<0.001	1.00	[1.00, 1.00]	<0.001
	Scopus H-Index comparison: without scholarship			*Scopus H-Index comparison: with scholarship*			*Scopus Citations comparison: without scholarship*			*Scopus Citations comparison: with scholarship*		
	Ratio effects	*[confint]*	*p*	*Ratio effects*	*[confint]*	*p*	*Ratio effects*	*[confint]*	*p*	*Ratio effects*	*[confint]*	*p*
Mean on controls	0.42	[0.33, 0.54]	<0.001	1.36	[1.28, 1.45]	<0.001	7.18	[4.62, 11.14]	<0.001	33.24	[28.14, 39.26]	<0.001
FAPESP Effect	4.27	[3.31, 5.50]	<0.001	1.41	[1.31, 1.51]	<0.001	6.54	[4.18, 10.24]	<0.001	1.42	[1.19, 1.70]	<0.001
Before PhD	1.26	[1.20, 1.32]	<0.001	1.18	[1.15, 1.21]	<0.001	1.00	[1.00, 1.01]	<0.001	1.00	[1.00, 1.00]	<0.001

Source: Research data.

The impact of scientific production measured through citations and the H-index is positive for FAPESP PhD scholarships, especially compared to individuals who had their PhD studies without funding support, using both Google Scholar and Scopus databases.

Considering the overall research performance of Brazilian researchers from 2010 to 2019 in Scopus, one can find 10,9 citations per publication regarding all knowledge fields. As expected, the ratio for students before completing their PhD is similar (1,00 citation per publication) and far below the Brazilian ratio for all groups. This implies that FAPESP scholarship has no effect during the project time, probably because journal publications are more common after research is ended and feedbacks from dissertation defense are incorporated.

However, FAPESP awardees, as well as other awardees, surpass the Brazilian average after this completion, revealing not only the influence of funding in the scientific production of PhD holders after the doctorate period has ended, but also the quality of research results from this group of investigators from São Paulo State, Brazil.

Figure 3 depicts the effect of the Google Scholar quote by knowledge field, comparing former FAPESP awardees with those who obtained funding from another agency and individuals who did not receive funding.

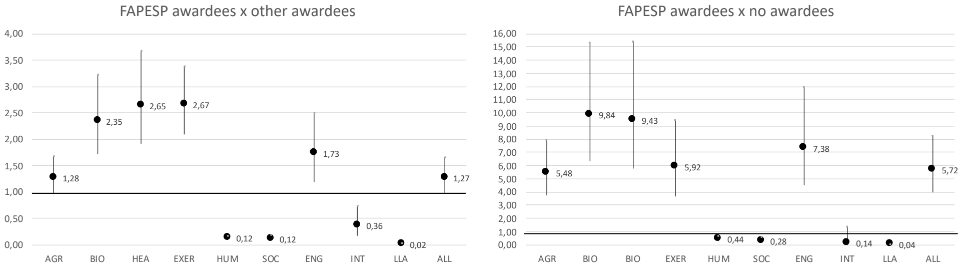


Figure 3. Google Scholar Citation Effect in Knowledge Fields

Source: Research data.

Figure 4, in turn, illustrates the effect of the H index on the knowledge field.

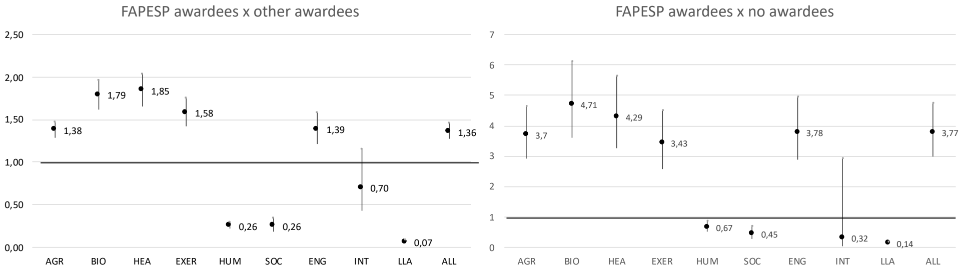


Figure 4. Google Scholar H Index Effect in Knowledge Fields

Source: Research data.

An analysis of the effects of citations and the H index by knowledge field (carried out for this report using only Google Scholar data, as stated above) confirms the generally positive effects of the FAPESP PhD scholarship, except for Humanities, Social Applied Sciences, Interdisciplinary Sciences and Linguistics, Languages and Arts, in which former scholarship holders from other agencies and those who took their PhD courses without a scholarship perform better than those from FAPESP. In other words, the pattern verified for the number of publications in journals previously analyzed is repeated, except for disciplines that have a different publication rationale not based on journal papers.

In short, our initial hypothesis is confirmed since the number of publications, the number of citations, and H index are higher for former FAPESP awardees when compared to individuals rejected by FAPESP. The effect – mainly shown after the PhD project has ended – is even more significant compared to those who did not have scholarships from other agencies, who took their PhD without funding support by a research agency. Through scholarships, the maintenance of these students allows them to focus entirely on their doctoral research and, consequently, to concentrate efforts to disseminate the achieved results.

Conclusions

This study's results allow us to state that the FAPESP PhD scholarship makes a quantitative and qualitative difference in scientific production. These effects are much more significant when comparing between former FAPESP scholarship holders and individuals who have completed a PhD course without a scholarship, revealing that despite the difference that the FAPESP scholarship makes, it is possible to state that PhD scholarships in the long run (either from FAPESP and other agencies) make much difference in the production and quality of scientific knowledge of PhDs, as already stated in other studies. Additionally, effects in scientific production quality (measured by citations and H-index) are detected in non-humanistic and non-social science knowledge areas, where journal papers are the preferred vehicle for disseminating research results. The impact of books and book chapters, although relevant, must be investigated with another kind of metric.

Therefore, it is also possible to state that the scholarships and what they mean in terms of dedication to graduate school represent fundamental conditions for knowledge production. Although it is a valid conclusion for the eligible study population (individuals who applied for FAPESP scholarships in the context of the State of São Paulo), the same likely happens in the broader Brazilian context, since the conditions for maintenance of researchers in other States are even more critical.

In this respect, the data is overwhelming: the quality of scientific production is firmly correlated to scholarships' granting. In this sense, if the goal of programs for supporting postgraduate studies through granting scholarships is to train scientists, it is possible to find much evidence of their success in the evaluation carried out.

A firm conclusion that emerges from the evaluation is that without scholarships (or another similar form of financing), there are essential threats to science in countries with Brazil's political and socioeconomic profile. This is because postgraduate students are one of the foundations of Brazilian (and so other countries') scientific and technological production and a renewing force of the advancement of knowledge made in public and private research centers and universities.

The immediate policy implication is the need to reinforce doctoral scholarship programs, which seems obvious to many developed countries that have reinforced support for education and scientific research in moments of crisis. However, it is controversial in the Brazilian context where governmental budgets decrease, and research funding agencies suffer from constant menaces.

Managerial implications concern the need for funding agencies to examine the behavior patterns related to different knowledge fields more deeply to establish appropriate mechanisms for support PhD students and evaluate the impacts of scholarships.

The main limitations of the study that we hope to address in the future are as follows: i) use information about peer review reports creating more comparable groups, such as best- rejected

applicants, as proposed by Bornmann et al. (2010) and Melin and Danell (2006) and worst-approved; ii) use gender as a new variable in this comparison of scientific production of PhDs; (iii) use coauthorship information to better understand the achievement of academic autonomy of these individuals and patterns of national and international collaboration.
Thus, we understand that there is a strong association between PhD scholarship and the quantity and quality of scientific production and that further investigation on this subject is still needed.

Acknowledgments

We gratefully acknowledge support from FAPESP under Award Number 2017/17461-1. The findings and results are solely the responsibility of the authors.

References

Ali, M. M., Bhattacharyya, P., & Olejniczak, A. J. (2010). The effects of scholarly productivity and institutional characteristics on the distribution of federal research grants. *The Journal of Higher Education*, 81(2), 164-178.

Andrade, R. D. O. (2019). Brazil budget cuts threaten 80,000 science scholarships. *Nature*, 572(7771), 575-576.

Angelo, C. (2017). Scientists plead with Brazilian government to restore funding. *Nature News*, 550(7675), 166.

Belavy, D. L., Owen, P. J., & Livingston, P. M. (2020). Do successful PhD outcomes reflect the research environment rather than academic ability? *PloS one*, 15(8), e0236327.

Benavente, J. M., Crespi, G., Garone, L. F., & Maffioli, A. (2012). The impact of national research funds: A regression discontinuity approach to the Chilean FONDECYT. *Research Policy*, 41(8), 1461-1475.

Bin, A., Salles-Filho, S. L., Capanema, L. M. Colugnati, F. (2015). What difference does it make? Impact of peer-reviewed scholarships on scientific production. *Scientometrics* (2015) 102: 1167. https://doi.org/10.1007/s11192-014-1462-9

Bin, A.; Salles-Filho, S. L.; Colugnati, F.; Campos, F. R. (2016). The ‘added value’ of researchers: the impact of doctorate holders on economic development. Auriol, L. et al. (2016), "*The Science and Technology Labor Force: The Value of Doctorate Holders and Development of Professional Careers*", Springer.

Böhmer, S., & Von Ins, M. (2009). Different—not just by label: research-oriented academic careers in Germany. *Research Evaluation*, 18(3), 177-184.

Bornmann, L., Leydesdorff, L., & Van den Besselaar, P. (2010). A meta-evaluation of scientific research proposals: Different ways of comparing rejected to awarded applications. *Journal of Informetrics*, 4(3), 211-220.

Escobar, H. (2019). Brazilian scientists lament ‘freeze’on research budget. *Science,* 364 (6436), p. 111

Goldsmith, S. S., Presley, J. B., & Cooley, E. A. (2002). *National Science Foundation Graduate Research Fellowship Program*, Final Evaluation Report. Virginia: NSF.

Halse, C.; Mowbray, S. (2011). The impact of the doctorate. *Studies in Higher Education*, 36(5), p. 513-525.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*,102(46), 16569-16572. http://dx.doi.org/10.1073/pnas.0507655102

Horta, H.; Cattaneo, M.; Meoli, M. (2018). PhD funding as a determinant of PhD and career research performance, *Studies in Higher Education*, 43:3, 542-570.

Jastrob, R.B.; Bukstein, D.; Güimil, X.U. (2012). *Informe de Evaluación: Impacto de becas de iniciación a la investigación - 2008*. Unidad de Evaluación y Monitoreo. Agencia Nacional de Investigación e Innovación. Documento de trabajo n.3.

Langfeldt, L., Bloch, C. W., Sivertsen, G. (2015). Options and limitations in measuring the impact of research grants—evidence from Denmark and Norway. *Research Evaluation*, 24(3), 256-270.

Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. *Scientometrics*, 90(2), 463-481.

Larivière, V. (2013). PhD students' excellence scholarships and their relationship with research productivity, scientific impact, and degree completion. *Canadian Journal of Higher Education*, 43(2), 27-41.

Leitch, S. (2006). *Prosperity for all in the global economy - world class skills*, Final Report, London: HMSO, 2006.

Lin, P. H., Chen, J. R., & Yang, C. H. (2014). Academic research resources and academic quality: a cross-country analysis. *Scientometrics*, 101(1), 109-123.

Lindahl, J., Colliander, C., & Danell, R. (2020). Early career performance and its correlation with gender and publication output during doctoral education. *Scientometrics*, 122(1), 309-330.

Melin, G., & Danell, R. (2006). The top eight percent: Development of approved and rejected applicants for a prestigious grant in Sweden. *Science and Public Policy*, 33(10), 702–712.

Meyer, N; Bührer, S. (2014). *Impact Evaluation of the Erwin Schrödinger Fellowships with Return Phase*. Final Report for the Austrian Science Fund (FWF), Vienna, Karlsruhe: Fraunhofer ISI.

Nelder, J.; Wedderburn, R. (1972). Generalized linear models. Journal of the Royal Statistical Society, 135(3), 370–384.

Neumann, R.; Tan, K.K. (2011). From PhD to initial employment: the doctorate in a knowledge economy. *Studies in Higher Education*, 36(5), p. 601–614.

Salter, A.J.; Martin, B.R. (2001). The economic benefits of publicly funded basic research: a critical review. *Research Policy*, 30, p. 509-532.

Stephan, P. E. (1996). The economics of science. *Journal of Economic literature*, 34(3), 1199-1235.

Tremblay, K. (2005). Academic Mobility and Immigration. *Journal of Studies in International Education*, 9(3), p. 196-228.

[i] Hirsch (2005) proposed that the H Index is the number of articles with citations greater than or equal to this number.

Gender differences in scientific careers: A large-scale bibliometric analysis

Hanjo Boekhout[1], Inge van der Weijden[2] and Ludo Waltman[2]

[1] *h.d.boekhout@liacs.leidenuniv.nl*
Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden (the Netherlands)

[2] *{i.c.m.van.der.weijden, waltmanlr}@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (the Netherlands)

Abstract

We present a large-scale bibliometric analysis of gender differences in science, covering all scientific disciplines and a large number of countries worldwide. We take a longitudinal perspective in which we trace the scientific careers of almost six million male and female researchers in the period 1996–2018. This enables us to analyze when male and female researchers entered the research system and how long they stayed in the system. It also allows us to study gender differences in productivity and seniority. Our analysis indicates that differences in the number of men and women entering the system are the most significant cause of gender imbalances. Differences between male and female researchers leaving the system do not play an important role. We also find that men produce substantially more publications than women and that men are more likely than women to move to more senior positions.

Introduction

Differences between men and women in their participation in scientific research are extensively being discussed. Underrepresentation of women in the research system may indicate that the system does not sufficiently benefit from the contributions that highly-qualified female researchers could make. To the extent that men and women differ in their research interests, it may also be a signal that the system does not provide sufficient room for female perspectives, causing certain topics, ideas, or approaches to be overlooked or marginalized.

In this paper, we present a large-scale bibliometric analysis of gender differences in science. Our analysis covers all scientific disciplines and a large number of countries worldwide. Rather than providing a cross-sectional picture of gender differences in the research system at a specific point in time, we offer a longitudinal perspective in which we trace the scientific careers of almost six million male and female researchers in the period 1996–2018. Such a longitudinal perspective enables us to analyze when male and female researchers entered the research system and how long they stayed in the system. It also allows us to study differences between male and female researchers in productivity (in terms of the number of publications they produce in specific years in their career) and seniority (in terms of the authorship positions they occupy in specific years in their career).

There is an extensive bibliometric literature on gender differences in science (for an overview, see Halevi, 2019). While earlier studies usually provided a cross-sectional picture of gender differences (e.g., Filardo et al., 2016; Holman et al., 2018; Jagsi et al., 2006; Larivière et al., 2013; West et al., 2013), more recent studies have started to offer longitudinal perspectives in which the careers of researchers are traced over time. A recent report by Elsevier (2020) for instance analyzed the development of male and female careers that started in 2009, and a recent study by Huang et al. (2020) presented a historical analysis by tracing male and female careers that ended between 1955 and 2010. Huang et al. concluded that gender differences in publication productivity and citation impact can largely be explained by differences in career length.

The importance of longitudinal analyses that account for factors such as career length and publication productivity is illustrated by research into gender differences in self-citations.

Based on an analysis of 1.5 million publications that appeared between 1779 and 2011, King et al. (2017) found that men self-cite 56% more than women. However, Mishra et al. (2018) showed that gender differences in self-citations largely disappear when accounting for a researcher's prior number of publications (for additional results pointing in the same direction, see Azoulay & Lynn, 2020), suggesting that men do not disproportionally self-cite compared with women.

Like the work by Huang et al. (2020), Mishra et al. (2018), and others, our work in the present paper offers a longitudinal bibliometric perspective on gender differences in science. By analyzing the careers of a very large number of researchers in all scientific disciplines, we aim to provide a solid high-level understanding of the relationship between gender and factors such as career start, career length, productivity, and seniority.

Data and methods

Bibliographic database

Our analysis is based on the Scopus database produced by Elsevier. Scopus bulk data is delivered annually by Elsevier to the Centre for Science and Technology Studies (CWTS) at Leiden University. To facilitate large-scale bibliometric analyses, the data is stored in a relational database at CWTS. Our analysis uses Scopus data received in April 2019. The data covers publications from 1996 to early 2019.

When two publications include the same author name, they may have been produced by the same author. However, they may also have been produced by two different authors that happen to have the same name. The other way around, when two publications do not include the same author name, they probably have not been produced by the same author. However, it may be that the publications have been produced by the same author, but the author may not have reported his or her name in a consistent way. An author name disambiguation algorithm attempts to determine which publications belong to the same author.

Our analysis relies on author identifiers provided by Scopus. These identifiers have been assigned using an author name disambiguation algorithm. In some cases, manual corrections have been made to the identifiers based on feedback from Scopus users. The author name disambiguation algorithm used by Scopus is a black box for which no public documentation is available. Based on an evaluation of 12,000 randomly selected authors, Scopus claims that its author identifiers have a precision of 98.1% and a recall of 94.4% (Baas et al., 2020). In an independent evaluation focused on Japanese authors, a precision and recall of 99.4% and 98.4%, respectively, were obtained (Kawashima & Tomizawa, 2015). For Dutch authors, lower values of 87.3% and 95.9% for precision and recall were found (Reijnhoudt et al., 2014). In a more recent evaluation focused on German authors, a precision and recall of 100.0% and 97.1% were reported (Aman, 2018). In the rest of this paper, we use the terms 'author' and 'researcher' to refer to authors as defined by author identifiers in Scopus.

Gender inference

For each author in Scopus, we attempted to algorithmically infer a gender, either male or female. If no gender could be inferred for an author, the gender was considered unknown. While we believe that algorithmic approaches to gender inference yield valuable insights, we emphasize that these approaches also have major limitations. In particular, algorithms for gender inference reduce gender to a binary concept, they are susceptible to bias, and their transparency is limited. We refer to Mihaljević et al. (2019) for a further discussion of these important issues.

Our algorithmic approach to gender inference uses three tools: Gender API (https://gender-api.com/), Gender Guesser (https://pypi.org/project/gender-guesser/), and Genderize.io (https://genderize.io/). Because of space limitations, we do not discuss our gender inference approach in more detail.

Gender statistics

In the calculation of the gender statistics reported in this paper, we considered only publications of the document types *article*, *chapter*, *conference paper*, *conference review*, and *review* in Scopus. Publications of other document types were disregarded. We considered only authors that have at least three publications. Authors with one or two publications are relatively likely to be the result of mistakes made by Scopus' author name disambiguation algorithm.

For each author, we determined the year of the first publication and the year of the last publication. We used the year of the first publication as a proxy of the year in which someone entered the research system. Likewise, the year of the last publication was used as a proxy of the year in which someone left the system. An important caveat applies. Because publications that appeared before 1996 are not included in our data, the year of someone's first publication could not always be correctly determined. The other way around, because someone may continue to publish in the future, the year of someone's last publication could not always be correctly determined either. For instance, someone's last publication in our data may be from 2017, but this person may have produced another publication in 2020. This person is then incorrectly considered to have left the research system in 2017. As a consequence of these issues, statistics for the first and last years covered by our analysis are less reliable than statistics for the years in between.

We also linked authors to scientific disciplines. We used the disciplinary classification provided by Scopus for this purpose. In this classification, each source (e.g., a journal or a conference proceedings) in Scopus belongs to one or more disciplines. The classification consists of two hierarchically related levels. We considered only the top level, which comprises 26 disciplines. We linked an author to a discipline if at least 80% of the publications of the author have appeared in sources belonging to the discipline. Some authors could not be linked to a discipline. Other authors were linked to multiple disciplines. (This is possible because a source may belong to more than one discipline.)

Finally, we linked authors to countries. An author was linked to a country if the author has an affiliation with an address in the country in at least 80% of his or her publications. Some authors could not be linked to a country. Other authors were linked to multiple countries. (This is possible because an author may have multiple affiliations in the same publication, possibly with addresses in different countries.) Our analysis covers only the 95 countries for which the share of authors for which the gender is unknown does not exceed 33%. For other countries, we do not consider gender statistics to be sufficiently informative. Our general statistics that do not pertain to a specific country cover all authors that have in at least 80% of their publications an affiliation with an address in one of the above-mentioned 95 countries.

Results

We first analyze gender differences for researchers entering and leaving the research system. We then consider gender differences in the productivity and seniority of researchers.

Our analysis includes 6.9 million researchers with at least three publications in the period 1996–2018. The analysis of researchers entering the system is based on 3.9 million male researchers and 2.1 million female researchers. In most of the results, the 0.9 million researchers for whom the gender could not be inferred are not included.

In the remaining analyses, we consider only the 205,779, 251,343, and 267,815 male and female researchers that entered the research system in, respectively, 2000, 2005, and 2010. By focusing on these three cohorts of researchers, we are able to separate gender differences in recent years from those in earlier years.

Entering the system

We first consider gender differences for researchers that enter the research system. Figure 1 shows the time trend in the percentage of men and women that entered the system. The left panel presents the time trend when researchers with an unknown gender are included in the statistics. The right panel presents the time trend when these researchers are excluded from the statistics. Like the statistics in the right panel, all remaining statistics reported in this paper do not include researchers with an unknown gender.

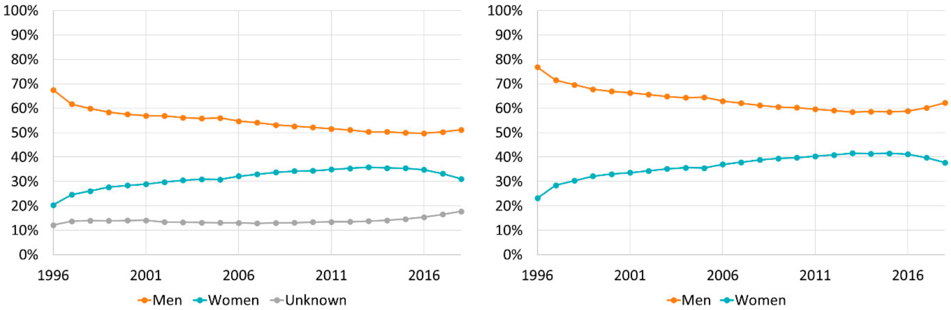


Figure 1. Percentage of men and women that entered the research system. Annual statistics for the period 1996–2018, including (left) and excluding (right) researchers for whom the gender is unknown.

Figure 1 shows a clear increase in the percentage of women that entered the system. Before 2000, for every two women that entered the system, there were four to five men that entered the system. In recent years, there were about three men that entered the system for every two women. We note that the statistics for 1996 and 1997 are not reliable because they also include many researchers that entered the system before 1996. Likewise, the statistics for the most recent years have a low reliability because some researchers that entered the system in these years do not yet have three publications and are therefore not included in the statistics. Most likely, the drop in the percentage of women that entered the system in recent years is an artefact resulting from the requirement of having at least three publications.

We also made a breakdown by discipline of the percentage of men and women that entered the system in 2000 and 2010 (not shown). In all disciplines in the physical sciences, engineering, and mathematics, the number of men that entered the system in 2010 was much larger than the number of women, ranging from 62% men in Chemistry to 85% men in Engineering. The strong underrepresentation of women in these disciplines is discussed extensively in the literature, including also in detailed bibliometric studies of specific disciplines, such as computer science (Cavero et al., 2015), mathematics (Mihaljević & Santamaría, 2020; Mihaljević-Brandt et al., 2016), and physics and astronomy (Mihaljević & Santamaría, 2020).

In most disciplines in the biomedical and health sciences, there were no major differences in the number of men and women that entered the system in 2010. In some disciplines, such as Environmental Science and Dentistry, the number of men that entered the system was somewhat larger than the number of women. In other disciplines, such as Immunology and Microbiology, the situation was the other way around. Nursing was the only exception. In this discipline, women represented 78% of all researchers that entered the system in 2010.

For disciplines in the social sciences and humanities, the picture is more mixed. The two extremes are Economics, Econometrics and Finance on the one hand, where women represented only 28% of all researchers that entered the system in 2010, and Psychology on the other hand, where women represented 66% of all researchers that entered the system. We refer to Lundberg and Stearns (2019) for an in-depth discussion of the underrepresentation of women in economics. In psychology, the overrepresentation of women among new researchers entering the system was also observed in a recent analysis by González-Alvarez and Sos-Peña (2020). When looking at researchers that entered in system in 2010, Arts and Humanities is close to gender parity. However, this may hide significant variation among different Arts and Humanities fields. Women are for instance reported to be strongly underrepresented in philosophy (Wilhelm et al., 2018).

For all disciplines except for Nursing, the percentage of women that entered the system in 2010 was higher than the percentage of women that entered the system in 2000. The difference is largest in Psychology, Environmental Science, and Chemical Engineering. In each of these disciplines, the percentage of women that entered the system increased by 12 percentage points between 2000 and 2010. Most disciplines moved in the direction of gender parity. However, Psychology is a notable exception (see also González-Alvarez & Sos-Peña, 2020). In this discipline, the percentage of women that entered the system increased from 54% in 2000 to 66% in 2010.

Figure 2 shows a breakdown by country of the percentage of women that entered the system in 2010. For countries that are colored gray, no statistics are available. Percentages below 30% can be found in some African countries, a number of countries in the Middle East, and also for Asian countries such as Japan (18%) and South Korea (21%). There are 12 countries where more than half of the researchers that entered the system in 2010 were female. In addition to a number of countries in Eastern and Southern Europe, this was the case for Puerto Rico (51%), Iceland (52%), Tunisia (56%), Philippines (57%), and Argentina (57%). We note that country-level statistics need to be interpreted with some care, because the accuracy of our approach to gender inference differs between countries.

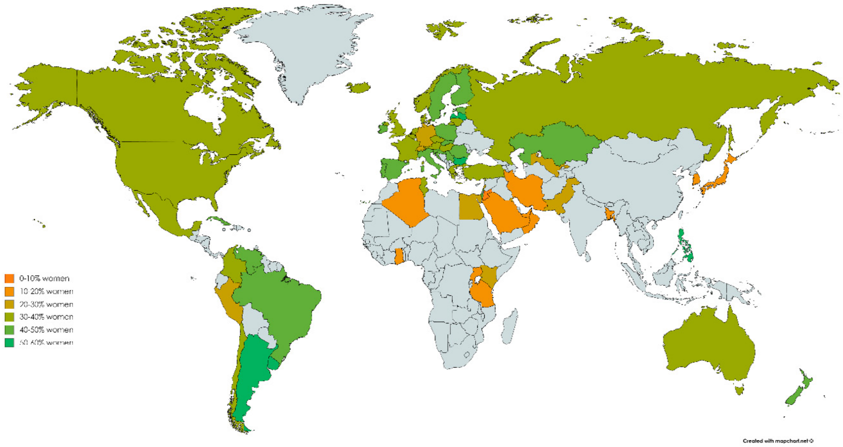


Figure 2. Percentage of women that entered the research system. Country statistics for researchers that entered the system in 2010.

Leaving the system

We now turn to gender differences for researchers that leave the research system. For male and female researchers that entered the system in 2000, 2005, and 2010, Figure 3 shows the annual percentage of researchers leaving the system. This percentage is calculated relative to all male and female researchers still in the system after a certain number of years. For instance, of the 137,715 male researchers that entered the system in 2000, 120,366 were still in the system after

5 years and 114,753 were still in the system after 6 years. Hence, after 5 years, (120,366 − 114,753) / 120,366 = 4.7% of the remaining male researchers left the system. This was the case for 5.2% of the remaining female researchers.

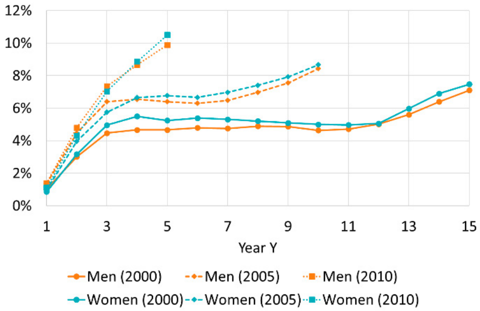


Figure 3. Percentage of men and women that left the research system after Y years relative to all men and women still in the system. Statistics for researchers that entered the system in 2000, 2005, and 2010.

For researchers that entered the system in 2005 or 2010, Figure 3 shows that in the early years of their career men were somewhat more likely to leave the system than women. In later years, women were more likely to leave the system. For researchers that entered the system in 2000, women consistently had a higher probability to leave the system than men (except for Y = 1 in Figure 3), but in later years of their career the difference was very small. We note that at the end of all three time trends in Figure 3 there is a substantial increase in the percentage of men and women leaving the system. This is an artefact resulting from researchers that are incorrectly considered to have left the system in recent years. These researchers have no publications in the most recent years covered by our analysis, but they will have publications in future years not covered by our analysis.

We also made a breakdown by discipline of the percentage of men and women leaving the system within five years after they entered the system (not shown). Statistics were calculated for researchers that entered the system in 2000 and 2010. For researchers entering in 2010, most disciplines show only a small difference between men and women. In some disciplines men were somewhat more likely to leave the system, while in other disciplines women had a somewhat higher probability to leave. The difference between men and women was more substantial in Nursing (44% vs. 35%) and Business, Management and Accounting (27% vs. 21%), with men being significantly more likely to leave the system than women in both disciplines. For researchers that entered the system in 2000, the situation was less balanced. In 19 of the 25 disciplines men were less likely to leave the system than women. The difference was largest in Psychology (13% vs. 20%) and Engineering (26% vs. 32%).

Results similar to ours were reported by Elsevier (2020) for researchers that entered the system in 2009. Huang et al. (2020) also performed a similar analysis, but their findings are different from ours. They found that each year female researchers are almost 20% more likely to leave the system than male researchers. The analysis of Huang et al. covers an older time period, which probably explains why their findings are different. For assistant professors hired since 1990 at US universities, it was found that both in science and engineering (Kaminski & Geisler, 2012) and in the social sciences (Box-Steffensmeier et al., 2015) there are no or almost no gender differences in retention rates. This seems to align with our finding that in later years of their career (i.e., after the PhD and postdoctoral period) female researchers that entered the system in 2000 were only slightly more likely to leave the system than their male colleagues.

Productivity

We next consider gender differences in the productivity of researchers. For male and female researchers that entered the research system in 2000, 2005, and 2010, Figure 4 shows the time trend in their average annual number of publications. The left panel presents statistics based on a full counting approach. In this approach, all co-authors of a publication are considered to have contributed to the publication, irrespective of the total number of co-authors of the publication. For all researchers that were still in the system in a certain year of their career, the left panel shows the average number of publications to which they contributed. The right panel presents statistics based on a fractional counting approach. In this approach, if a publication is co-authored by *n* researchers, each of these researchers is considered to have produced (1 / *n*)th of a publication. For all researchers that were still in the system in a certain year of their career, the right panel shows the average number of fractionally counted publications they produced.

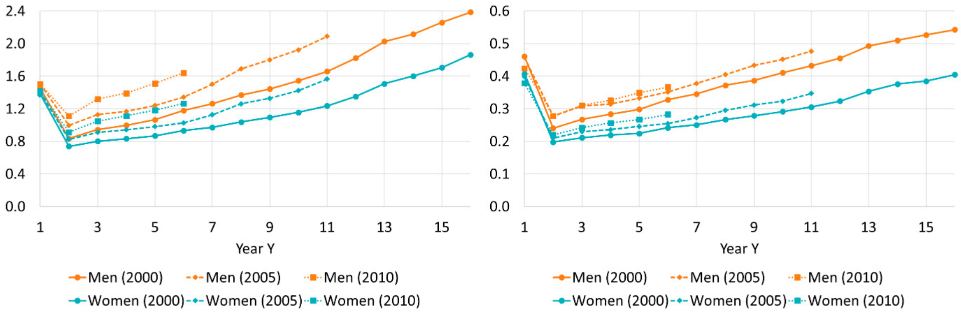


Figure 4. Average number of publications of men and women in year Y of their scientific career. Statistics for researchers that entered the research system in 2000, 2005, and 2010 based on full (left) and fractional (right) counting.

Figure 4 shows that, regardless of the year in which they entered the system and the year of their career, men consistently had a substantially higher productivity than women. Based on the full counting statistics, the productivity of men was between 20% and 35% higher than the productivity of women. The difference is even larger when looking at the fractional counting statistics. Based on these statistics, the productivity of men was between 25% and 40% higher than the productivity of women.

There are well-known differences between disciplines in the average number of publications of a researcher. The gender differences observed in Figure 4 could be due to a relative overrepresentation of men in disciplines with a larger average number of publications per researcher and, conversely, a relative overrepresentation of women in disciplines with a smaller average number of publications per researcher. To explore this possibility, we made a breakdown by discipline of the average number of publications of men and women in year 6 of their career (not shown). Statistics were calculated for researchers that entered the system in 2000 and 2010, based on both a full counting approach and a fractional counting approach.

We first consider the full counting statistics. For researchers that entered the system in 2000, in year 6 of their career male researchers had a higher productivity than their female colleagues in 22 of the 25 disciplines. The same observation can be made for researchers that entered in 2010. On average, the productivity per discipline was 17% higher for male researchers that entered in 2010 than for their female colleagues (calculated by averaging the percentage difference over all disciplines, weighing each discipline by its number of researchers). Hence, the overall gender difference in productivity observed in the left panel of Figure 4 can also be found at the level of individual disciplines. However, at the disciplinary level, the relative differences between men and women were somewhat smaller than at the overall level. This means that the overall gender difference in productivity was partly caused by a relative overrepresentation of men in disciplines with a larger average number of publications per researcher. In particular, the average number of publications per researcher was much larger in Physics and Astronomy than in other disciplines, and Physics and Astronomy was also one of the disciplines with the largest overrepresentation of men among researchers entering the system.

Like the full counting statistics, the fractional counting statistics indicate that in most disciplines men were more productive than women in year 6 of their career. The fractional counting approach yields somewhat larger relative differences between men and women than the full counting approach. For researchers that entered the system in 2010, on average the productivity per discipline was 20% higher for male researchers than for their female colleagues. Based on the fractional counting approach, the average number of publications per researcher was substantially larger in disciplines in the physical sciences, engineering, and mathematics than in disciplines in the biomedical and health sciences. The overrepresentation of men among researchers entering the system was also much larger in the physical sciences, engineering, and mathematics than in the biomedical and health sciences. This means that the overall gender difference in productivity observed in the right panel of Figure 4 was caused partly by gender differences in productivity at the level of individual disciplines and partly by the relative overrepresentation of men in certain disciplines, in particular in the physical sciences, engineering, and mathematics.

Productivity differences between male and female researchers, with male researchers on average producing more publications than their female colleagues, have been found in a large number of studies (for an overview, see Halevi, 2019). These differences, which lead to the so-called productivity puzzle (Cole & Zuckerman, 1984), were also observed in recent large-scale analyses by Elsevier (2020) and Mishra et al. (2018). In line with our results reported above, the literature generally finds that productivity differences cannot be explained by differences in career length or career stage. However, a notable exception is a recent study by Huang et al. (2020). Based on a large-scale longitudinal analysis, Huang et al. found that differences in career length explain differences between male and female researchers in productivity, leading to the conclusion that "men and women publish at a comparable annual rate", referred to as a 'gender invariant' by Huang et al. These findings are in stark contrast with our findings presented above as well as the findings of many other studies. The discrepancy may be due to the older time period covered by the analysis of Huang et al. Moreover, unlike our analysis, the analysis of Huang et al. does not consider the productivity of researchers in a specific period in their career, but only their average productivity during their entire career. This may also explain why the findings of Huang et al. are different.

Seniority

Finally, we look at gender differences in the seniority of researchers. We use a researcher's position in the list of authors of a publication as a crude proxy of seniority. More specifically, we focus on the first and last author of a publication. In most disciplines, being first author of a publication indicates that one has made a major contribution to a research project, either in a more junior or in a more senior role. Being last author of a publication is especially relevant in biomedical disciplines, where the last author typically is a senior researcher that carries the overall responsibility for a research project. Although statistics on first and last authorship provide useful information, we emphasize that there are important differences between disciplines in the way in which the order of the authors of a publication is determined. In some disciplines, in particular in economics, high energy physics, and mathematics, it is quite common to list authors in alphabetical order (Waltman, 2012). First and last authorship have little or no meaning in these disciplines. We refer to Marušić et al. (2011) for an overview of the literature on authorship practices.

For earlier large-scale studies of gender differences in first and last authorship, we refer to Holman et al. (2018) and West et al. (2013). There are also several studies with a specific focus on medical journals (e.g., Filardo et al., 2016; Jagsi et al., 2006). Importantly, unlike our analysis reported below, earlier studies were of a cross-sectional nature. These studies therefore could not account for factors such as the career length of researchers.

For male and female researchers that entered the research system in 2000, 2005, and 2010, Figure 5 shows the time trend in the probability of being first (left panel) or last (right panel) author of a publication. In the early years of their career, researchers were relatively likely to be first author of a publication. In later years, the probability of being first author decreased and researchers were more likely to be last author. However, there are significant gender differences. As can be seen in the left panel of Figure 5, in the early years of their career, men were more likely than women to be first author of a publication. The opposite pattern can be observed for later years. Moreover, as shown in the right panel of Figure 5, both in earlier and in later years of their career, men were substantially more likely to be last author than women. The probability of being last author was between 15% and 30% higher for men than for women. Since last authorship is an indication of seniority, especially in biomedical disciplines, this suggests that on average it took less time for men than for women to reach a more senior position in the research system.

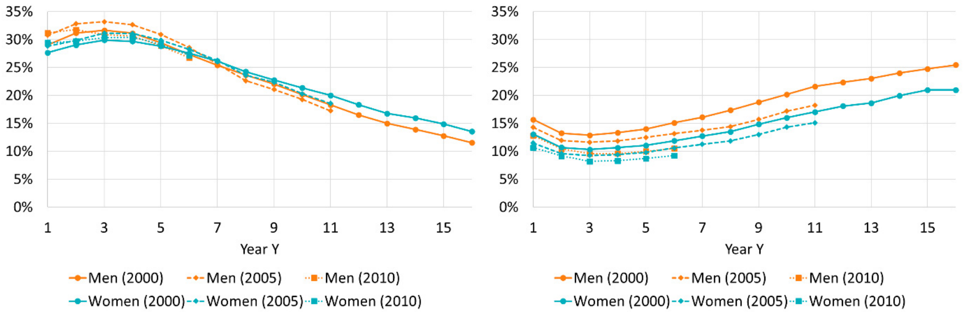


Figure 5. Probability of being first (left) or last (right) author of a publication for men and women in year Y of their scientific career. Statistics for researchers that entered the research system in 2000, 2005, and 2010.

As already mentioned, different disciplines have different norms for determining the order of the authors of a publication. For men and women in year 6 of their career, we made a breakdown by discipline of the probability of being last author of a publication (not shown). The statistics pertain to researchers that entered the system in 2000 and 2010.

From the viewpoint of understanding gender differences in researchers' seniority, the most interesting results are the statistics for last authorship in the biomedical and health sciences. Looking at researchers that entered the system in 2010, we find that in all disciplines in the biomedical and health sciences men had a higher probability than women to be last author of a publication in year 6 of their career. For researchers that entered in 2000, this was the case in 9 of the 11 disciplines in the biomedical and health sciences. In terms of the number of researchers, Medicine is by far the largest discipline in the biomedical and health sciences, followed by Biochemistry, Genetics and Molecular Biology and Agricultural and Biological Sciences. In each of these disciplines, for researchers that entered the system in 2010 and that were in year 6 of their career, the probability to be last author of a publication was about 25%

higher for men than for women, suggesting that on average men moved to more senior roles more quickly than women. Underrepresentation of women among last authors was also observed by Holman et al. (2018) and West et al. (2013), but these cross-sectional studies did not account for differences in career length and other factors.

Discussion and conclusion

The large-scale bibliometric analysis presented in this paper provides detailed insights into differences between men and women in entering and leaving the research system and in their productivity and seniority while they are in the system.

The percentage of women entering the research system shows a clear increasing trend. About 40% of all researchers that entered the system in recent years were female. In 2000 this was the case for only 33% of the researchers, and before 2000 the percentage of women entering the system was even lower. There are large differences between disciplines. In the physical sciences, engineering, and mathematics, only 22% of the researchers that entered the system in 2010 were female. In economics, this was the case for just 28% of the researchers. In contrast, in nursing and psychology, women represented, respectively, 78% and 66% of the researchers that entered the system in 2010. Differences between countries are large as well. In some countries, for instance in Eastern and Southern Europe, more than half of the researchers that entered the system in 2010 were female. In other countries, such as Japan and South Korea, this was the case for only one out of every five researchers.

Women seem to be somewhat more likely to leave the research system than men. This difference is most clearly visible for researchers that entered the system in 2000. For researchers that entered in 2005 and 2010, the difference is less clear. In the first few years of the careers of these researchers, women were actually less likely to leave the system than men. In later years, however, women had a higher probability to leave than men. Even for researchers that entered in 2000, the gender difference in leaving the system was not very large. As mentioned above, 33% of the researchers that entered in 2000 were female. After 15 years, 52% of the male researchers and 54% of the female researchers had left the system, resulting in 32% of the researchers still in the system being female.

There are substantial gender differences in productivity, where we define productivity in terms of the number of publications produced by a researcher in a certain year of his or her career. Depending on the year in which they entered the research system, the year of their career, and the counting approach (i.e., full or fractional counting), men were on average between 20% and 40% more productive than women. This can partly be explained by the strong overrepresentation of men in disciplines with a larger average number of publications per researcher, in particular in the physical sciences, engineering, and mathematics. After correcting for this overrepresentation, we found that in year 6 of their career men were on average between 15% and 20% more productive than women.

Determining researchers' seniority is challenging, but being last author of a publication can be used as crude proxy, especially in biomedical disciplines. Gender differences in last authorship are substantial. Overall, men were between 15% and 30% more likely to be last author of a publication than women. For researchers in biomedical disciplines that entered the research system in 2010 and that were in year 6 of their career, the probability of being last author was about 25% higher for male researchers than for their female colleagues. These statistics suggest that men moved to more senior roles more quickly than women.

Based on the various factors considered in our analysis, the overall picture emerging from our work is that differences in the number of men and women entering the research system are the most significant cause of gender imbalances in the system. In recent years, the number of male researchers entering the system was between 40% and 50% larger than the number of female researchers entering the system. In earlier years, the difference was even larger, and the effect of this is still visible in the current composition of the scientific workforce. It will be interesting to see whether the increasing trend in the proportion of women entering the research system will continue in the coming years. Compared with the gender differences that we found for researchers entering the system, differences between male and female researchers leaving the system are small and do not play an important role in explaining overall gender imbalances in the system. However, our statistics on last authorship do suggest that within the research system men and women follow significantly different career trajectories, with men being more likely than women to move to more senior positions. Finally, publications play an important role in the research system, which leads to an additional gender imbalance, since men produce on average between 15% and 20% more publications than women.

We end this paper by drawing attention to a number of important limitations of our research. First of all, our work is subject to the limitations of the Scopus database. The coverage of the scientific literature provided by the Scopus database is incomplete. Some researchers therefore are not included in our analysis, while for other researchers their publication oeuvre is covered only partly. In addition, the Scopus data to which we have access does not go back before 1996. This means that our analysis does not properly cover researchers with long careers in science. The disciplinary classification provided by Scopus also has weaknesses, such as questionable classifications of some journals and large size differences between disciplines. Furthermore, our analysis relies on algorithmic approaches to author name disambiguation and gender inference. These approaches inevitably make mistakes that affect our analysis. For instance, Scopus' author name disambiguation algorithm may sometimes fail to recognize that two authors in fact represent the same person. Our approach to gender inference has the limitation of relying on a binary notion of gender. Moreover, for many researchers our approach is unable to infer a gender, leading to missing statistics for important countries such as China and India. Another limitation of our research is that all publications are treated in the same way. The citation impact of publications for instance is not taken into account. Finally, while a large-scale bibliometric analysis like the one presented in this paper offers valuable high-level insights into gender differences in science, it does not provide a more in-depth understanding of gender differences in specific disciplines or countries and of the underlying mechanisms causing these differences. To obtain such an understanding, our work needs to be complemented with case studies in which gender differences in individual disciplines or countries are analyzed in more detail.

References

Aman, V. (2018). Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. *Scientometrics*, 117(2), 705–720. https://doi.org/10.1007/s11192-018-2895-3

Azoulay, P. & Lynn, F.B. (2020). Self-citation, cumulative advantage, and gender inequality in science. *Sociological Science*, 7, 152–186. https://doi.org/10.15195/v7.a7

Baas, J., Schotten, M., Plume, A., Côté, G. & Karimi, R. (2020). Scopus as a curated, high quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, 1(1), 377–386. https://doi.org/10.1162/qss_a_00019

Box-Steffensmeier, J.M., Cunha, R.C., Varbanov, R.A., Hoh, Y.S., Knisley, M.L. & Holmes, M.A. (2015). Survival analysis of faculty retention and promotion in the social sciences by gender. *PLOS ONE*, 10(11), e0143093. https://doi.org/10.1371/journal.pone.0143093

Cavero, J.M., Vela, B., Cáceres, P., Cuesta, C. & Sierra-Alonso, A. (2015). The evolution of female authorship in computing research. *Scientometrics*, 103(1), 85–100. https://doi.org/10.1007/s11192-014-1520-3

Cole, J.R. & Zuckerman, H. (1984). The productivity puzzle: Persistence and change in patterns of publication of men and women scientists. In M.W. Steinkamp & M.L. Maehr (Eds.), *Advances in Motivation and Achievement: Women in Science* (pp. 217–258). JAI Press.

Elsevier (2020). *The researcher journey through a gender lens*. Elsevier. https://www.elsevier.com/research-intelligence/resource-library/gender-report-2020

Filardo, G., da Graca, B., Sass, D.M., Pollock, B.D., Smith, E.B. & Martinez, M.A.M. (2016). Trends and comparison of female first authorship in high impact medical journals: Observational study (1994–2014). *BMJ*, 352, i847. https://doi.org/10.1136/bmj.i847

González-Alvarez, J. & Sos-Peña, R. (2020). Women publishing in American Psychological Association journals: A gender analysis of six decades. *Psychological Reports*, 123(6), 2441–2458. https://doi.org/10.1177/0033294119860257

Halevi, G. (2019). Bibliometric studies on gender disparities in science. In W. Glänzel et al. (Eds.), *Handbook of science and technology indicators* (pp. 563–580). Springer. https://doi.org/10.1007/978-3-030-02511-3_21

Holman, L., Stuart-Fox, D. & Hauser, C.E. (2018). The gender gap in science: How long until women are equally represented? *PLOS Biology*, 16(4), e2004956. https://doi.org/10.1371/journal.pbio.2004956

Huang, J., Gates, A.J., Sinatra, R. & Barabási, A.L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences of the USA*, 117(9), 4609–4616. https://doi.org/10.1073/pnas.1914221117

Jagsi, R., Guancial, E.A., Worobey, C.C., Henault, L.E., Chang, Y., Starr, R., ... & Hylek, E.M. (2006). The "gender gap" in authorship of academic medical literature—a 35-year perspective. *New England Journal of Medicine*, 355(3), 281–287. https://doi.org/10.1056/NEJMsa053910

Kaminski, D. & Geisler, C. (2012). Survival analysis of faculty retention in science and engineering by gender. *Science*, 335(6070), 864–866. https://doi.org/10.1126/science.1214844

Kawashima, H. & Tomizawa, H. (2015). Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics*, 103(3), 1061–1071. https://doi.org/10.1007/s11192-015-1580-z

King, M.M., Bergstrom, C.T., Correll, S.J., Jacquet, J. & West, J.D. (2017). Men set their own cites high: Gender and self-citation across fields and over time. *Socius*, 3. https://doi.org/10.1177%2F2378023117738903

Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C.R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, 504, 211. https://doi.org/10.1038/504211a

Lundberg, S. & Stearns, J. (2019). Women in economics: Stalled progress. *Journal of Economic Perspectives*, 33(1), 3–22. https://doi.org/10.1257/jep.33.1.3

Marušić, A., Bošnjak, L. & Jerončić, A. (2011). A systematic review of research on the meaning, ethics and practices of authorship across scholarly disciplines. *PLOS ONE*, 6(9), e23477. https://doi.org/10.1371/journal.pone.0023477

Mihaljević, H., & Santamaría, L. (2020). Authorship in top-ranked mathematical and physical journals: Role of gender on self-perceptions and bibliographic evidence. *Quantitative Science Studies*, *1*(4), 1468–1492. https://doi.org/10.1162/qss_a_00090

Mihaljević, H., Tullney, M., Santamaría, L. & Steinfeldt, C. (2019). Reflections on gender analyses of bibliographic corpora. *Frontiers in Big Data*, 2, 29. https://doi.org/10.3389/fdata.2019.00029

Mihaljević-Brandt, H., Santamaría, L. & Tullney, M. (2016). The effect of gender in the publication patterns in mathematics. *PLOS ONE*, 11(10), e0165367. https://doi.org/10.1371/journal.pone.0165367

Mishra, S., Fegley, B.D., Diesner, J. & Torvik, V.I. (2018). Self-citation is the hallmark of productive authors, of any gender. *PLOS ONE*, 13(9), e0195773. https://doi.org/10.1371/journal.pone.0195773

Reijnhoudt, L., Costas, R., Noyons, E., Börner, K. & Scharnhorst, A. (2014). 'Seed + expand': A general methodology for detecting publication oeuvres of individual researchers. *Scientometrics*, 101(2), 1403–1417. https://doi.org/10.1007/s11192-014-1256-0

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics*, 6(4), 700–711. https://doi.org/10.1016/j.joi.2012.07.008

West, J.D., Jacquet, J., King, M.M., Correll, S.J. & Bergstrom, C.T. (2013). The role of gender in scholarly authorship. *PLOS ONE*, 8(7), e66212. https://doi.org/10.1371/journal.pone.0066212

Wilhelm, I., Conklin, S.L. & Hassoun, N. (2018). New data on the representation of women in philosophy journals: 2004–2015. *Philosophical Studies*, 175(6), 1441–1464. https://doi.org/10.1007/s11098-017-0919-0

How to do research on the societal impact of research? Studies from a semantic perspective

Andrea Bonaccorsi[1], Filippo Chiarello[1], Nicola Melluso[1] and Gualtiero Fantoni[2]

[1] *a.bonaccorsi@gmail.com, nicolamelluso@gmail.com, filippo.chiarello@unipi.it*
Department of Energy, Systems, Territory and Construction Engineering, University of Pisa, Pisa (Italy)

[2] *gualtiero.fantoni@unipi.it*
Department of Civil and Industrial Engineering, University of Pisa, Pisa (Italy)

Abstract

We review some recent works of our research lab that have applied novel text mining techniques to the issue of research impact assessment. The techniques are Semantic Hypergraphs and Lexicon-based Named Entity Recognition. By using these techniques, we address two distinct and open issues in research impact assessment: the epistemological and logical status of impact assessment, and the construction of quantitative indicators.

Introduction

In the latest few months, we have published four papers in various journals that contribute to the debate on research impact (Bonaccorsi et al. 2020; 2021a; 2021b; 2021c). The overall aim of this effort is to call the attention on two dimensions of the research impact that have been somewhat neglected. The first is a theoretical issue: What kind of statement is a statement of the type "research X has produced an impact on Y"? What is the logical nature of this statement? What is the semantic structure of a statement of this type? It turns out that addressing this apparently theoretical question has far reaching implications. The second is a methodological issue: Which methods can we use for research impact assessment? Can we build indicators, as it happens in other areas of S&T? In this paper we deepen the conceptual issues underlying these papers, we describe their findings, and illustrate a research program for future studies and practical applications.

The historical nature of research impact statements

What kind of statement is a statement of the type "research X has produced an impact on Y"? When asked to give account of the impact of their research, researchers make use of a narrative. This is dictated by the nature of the requirement: the impact of research is a process that unfolds over time, involving several actors, with many events, facts, accidents taking place at different points in time. The outcome is a change of state that does not take place abruptly but is prepared by a sequence of events and a plurality of actors. To make sense of the unfolding of the impact process the only type of writing is a narrative one. Researchers must persuade their readers that their research has indeed taken part to a historical process, whose reality can be demonstrated, in which some actors have directly or indirectly benefited from it, leading to some improvement. To persuade the readers the researchers must do two things: first, it must build a narrative that unfolds over time, is plausible and realistic, and ends up with the impact; second, it must demonstrate that within this narrative, researchers have had a role, that is, have produced the impact, or have contributed to the production of the impact.

The narrative style is shared by two fundamentally different types of writing: history and fiction. Historical narrative is expected to reconstruct sequences that represent historical facts, or events that have taken place in historical time, according to the best available documentation. Historical narratives may have a subjective flavour in the way in which the flow of events is reconstructed and the importance is weighted, but they must comply with severe standards of control about the historical truthfulness of their statements. In contrast, fiction has no obligation whatsoever with respect to truth.

In order to build a flow of events that make sense, when writing a narrative text authors place of attention in giving details, offering a rich and contextualized description of what indeed happened. Vividness, detail, and richness of the description are essential component of narratives.

The narrative nature of research impact reports, however, may reasonably lead to the claim that they cannot include causal statements. To have a causal statement one must satisfy some general requirements that cannot be satisfied if the action takes place in the past. According to an influential tradition of research in post-structuralism, if the events narrated are placed in the past, then there is no difference between history and fiction, because the readers can never control the objectivity of the events.

Causality and credibility

In our first paper (Bonaccorsi et al. 2021a) we argue that research impact reports do include propositions that have a causal value. We defend this argument in two ways: making reference to the philosophical debate on the nature of historical knowledge, and by applying to impact statements a new text mining technique that makes it possible to identify the semantic structure of complex text structures, such as entire sentences, called Semantic Hypergraphs. By applying this technique to the collection of REF impact case studies we achieve some interesting results.

If we agree that research impact statements are historical statements, then we can move forward and ask some questions that have been the object of a passionate debate in the ‘50s and ‘60s, which has been renewed in recent times. The questions can be formulated as follows: Do historical narratives include explanatory statements? Can historical explanation be considered a scientific explanation? If the answer is positive, under which formal conditions do historical statements have explanatory value? If the answer is negative, how can historians aim at scientific objectivity?

The debate was sparked by the celebrated article by Carl Hempel. The argument was sharp: for a statement to be explanatory, it must include reference to a law-like proposition and a specific, contextualized event, given some boundary conditions. The explanation is the logical process by which we demonstrate that the specific event (explanandum) is logically entailed by the general law (explanans), given the boundary conditions. We perform this task in a logical way, by subsuming the individual event into a general category, following a nomological-deductive approach.

This formulation had an enormous impact. We might say it is, explicitly or implicitly, at the root of all arguments that sharply separate between hard science and humanities, negating the scientific value of the latter. Interestingly, it has been rejected with a number of strong arguments by philosophers and historians. We summarize these arguments here, before showing why they are relevant to the debate on research impact.

First of all, the nomological-deductive model of explanation is not the only possible model of causality. It is a model of necessary and sufficient conditions, which requires for any explanation the existence of general laws that treat the specific event as a member of a class. In other words, it requires the existence of a class of entities. The class of entities is demonstrated by repetition, that is, by producing experimentally manipulated pieces of evidence that reproduce with regularity the same outcomes.

This is not the only logically valid type of explanation. Another logically valid type is a statement which claims that a given condition X has been partially sufficient for Y. This statement would not treat X as a case in a class of entities that can be reproduced. By reasoning on X historians can build up a collection of partial explanations which together produce a sufficient reason for the manifestation of Y. Each of them has local and circumstantial value.

Second, historians do make use of general laws in their formulations. As stated by Nicholas Rescher, historians are consumers of general laws, not producers. They do not ignore the physical or chemical general laws that dominate the working of nature and they clearly see their application in specific cases. It is rare, however, that the implication of a general law is the crucial element to be invoked to formulate a historical judgment. As the historian Marc Bloch stated, it is certainly true that it is the law of gravity that governs the fact that King X fell from his horse and died. But it is not so interesting to know. Other less general explanatory factors, such as the speed at which he was riding or the threat he was addressing are more important. As the philosopher noted, such general laws very often do not exist, or are trivial.

Third, the validity of historical statements is not predicated on the logical strength of a deductive reasoning, but on the completeness and accuracy of historical documentation, and the plausibility of the narrative reconstruction of causal linkages. It is certainly true that it is not possible to validate experimentally all causal linkages between events. However, historians collect all possible causal explanations (including those that are generated by true general laws) and select among them those that can create a plausible chain of events. In doing so they make appeal to the largest available documentation and to patterns of reasoning that are not deductive but abductive, following the "method of clues" illustrated by Carlo Ginzburg.

On the causality value of research impact statements

From this purely theoretical discussion we move towards another question: if research impact statements have historical nature, what kind of explanation do they include?

Here it is important to call into the debate the influential line of thinking that has proposed the notion of contribution, as opposed to attribution, as the logical foundation for impact assessment. According to this argument, it is impossible to control for all potentially influencing factors that may lead to a research impact. The social impact of research takes place within complex and multidimensional processes that extended over long, often unpredictable, time horizons.

Consequently, the notion of attribution, or the process by which a specific event may be logically demonstrated to be dependent upon a specific condition, must be rejected, in favour of a weaker notion of contribution. We agree with this argument with a qualification. It is one thing to reject the notion of attribution if we assume that the only causal model underlying the attribution is the nomological-deductive causality suggested by Hempel, or variants of this model that fit the way in which causality is assumed in hard sciences. It is another thing to suggest that impact statements do not include any kind of causality. We show that they indeed include causal statements, but these statements must be understood in the light of a theory of historical, not nomological, causality.

In order to examine this issue we adopt a recently developed technique in text mining called Semantic Hypergraphs (Menezes and Roth, 2019). This technique overcomes one of the most important limitation of Natural Language Processing techniques, that is, the ability to examine individual word, or short sequences of words (n-grams), due to computational limits of the algorithms. With Semantic Hypergraphs the meaning of the text is not reconstructed via the statistical analysis of frequency and clustering of words, but by the construction of higher-level topological structures in which the meaning of words is derived from their relational position with all other words. The unit of analysis is a sentence, that is a chunk of text included between two periods. The authors develop a systematic language that allows the automatic processing of sentences. We have applied this technique to the REF collection of impact case studies, following the flowchart in Figure 1 (source: Bonaccorsi et al. 2021a).

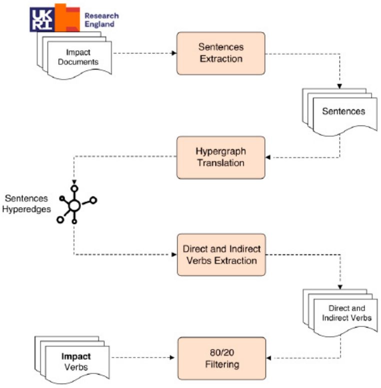


Figure 1. Workflow for the extraction of impact verbs from the REF corpus.

By looking at the structure of sentences we have identified the following elements
i) Direct impact verb
ii) Indirect impact verb
iii) Agent
iv) Topic
The technique allows a clear distinction between impact verbs and non-impact verbs. Impact verbs imply an action of something over something else. This structure is systematically discovered in sentences included in the REF reports. Impact verbs can be direct or indirect, depending on their position in the sentence (technically speaking, the rank in the depth of sentences). On the basis of this distinction and of statistical definitions, it is possible to clearly identify impact sentences and non-impact sentences. At this point we have a powerful tool to ask questions about the distribution of impact vs non-impact sentences across the collection of REF documents. We build up an indicator (IR) as the ratio between impact and non-impact sentences. We find a number of interesting insights:

- Impact sentences have a clear causal structure: they try to demonstrate that X has produced a change in Y
- To achieve the goal of showing the causal effect there is a preparation, mostly of descriptive type, that is done by means of non-impact sentences
- Reports in SSH make larger use of non-impact sentences, meaning that they need to establish a larger and more detailed descriptive evidence before claiming for impact
- The verbal structure of reports, that is, the pattern of utilization of direct and indirect verbs, is not different between SSH and STEM

We interpret these findings as suggesting that research impact statements do have a causal structure, with no difference between SSH and STEM. At the same time we find an interesting difference, in the sense that the construction of the causal statements is longer, more articulated, more complex in SSH reports.

We deepen this issue by introducing the notion of credibility of historical statements in our second paper (Bonaccorsi et al. 2021b). While for professional historians we may assume that rigorous methodological standards are adopted, as it is clear from the methodological and epistemological debate discussed above, the same cannot be said for REF reports. They are written by researchers, or university administrators, or consultants. They have inevitably a rhetorical and instrumental value: the must persuade the evaluators, call their attention to the importance of the impact, and obtain a high score. Remember that within the framework of REF, no less than 20% of funding is allocated on the basis of the impact assessment.

Nevertheless, the authors of the REF reports clearly understand that all these instrumental and practical goals cannot be achieved at all if the statements are not credible. Since they are placed in the past, the readers do not have any experimental control on them. But they mentally reason about the plausibility of the narrative reconstruction. By knowing the pragmatic orientation of the authors, the readers must challenge the credibility of the arguments.

We examine this issue by applying to the causal argumentation the criteria for historical explanation proposed by the philosopher Carl Hammer (2008). According to this philosopher, historical statements have causal power if they include "partial sufficient conditions that are normative, identifiable, manipulable, and not easily replaceable" (Hammer, 2008, 198). We refer the readers to our paper for a full scale, but non-technical discussion of these criteria. By applying these criteria to the main areas of SSH research (social sciences, humanities) and STEM (medicine, technology) and to their main fields of social impact, we derive a kind of theory of credibility of research impact statements. According to this theory, the aim of credibility is more difficult in SSH than in STEM. This is because STEM causality statements may make reference to an established repository of highly structured and formalized processes, in some cases standardized in the public regulation or business practice. During the pathways of impact, the achievement of causal effects is witnessed by the production of formal and socially identified intermediate outputs (e.g., clinical trial, patent, prototype). The description of these chains by the authors of REF reports activates in the mind of evaluators a pattern of recognizability and familiarity. The causal linkages suggested at each of the junctures of the pathway are highly credible. This is not the case for SSH. Here the chain is longer and more fragile. We elaborate on this issue in the final sections of this paper.

Mapping users of research

In two other papers (Bonaccorsi et al. 2020; 2021c) we make use of another technique from machine learning, i.e., Named Entity Recognition. We identify all names in REF documents that refer to social groups, or groups that may be the intended or unintended, direct or indirect beneficiaries of the research of universities. These groups are mentioned in a variety of ways in REF reports. We use a lexicon-based approach, by filtering the REF texts with a lexicon of users with 76,857 entries, developed after extensive research.

With the help of this lexicon, it is possible to saturate the semantic space of research beneficiaries, identifying all social groups that are implied in the impact process. After the extraction we are able to obtain two applications. The first is a mapping exercise, in the tradition of science maps. We draw the entire map of users of research of UK universities, modulating the granularity of the representation. We then cluster the user groups by using community detection techniques. A very informative map is reproduced in Figure 2 (source: Bonaccorsi et al. 2020). All identified clusters are consistent and well delineated. By modulating the granularity we can zoom further in the representation.

The second application is the construction of indicators. Since our lexicon saturates the semantic space, we are in a position to define indicators with appropriate statistical properties. We define the following indicators:

- Frequency
- Diversity
- Specificity

We run the calculation on the entire REF collection and then separately by discriminating between SSH and STEM. This offers some interesting insights.

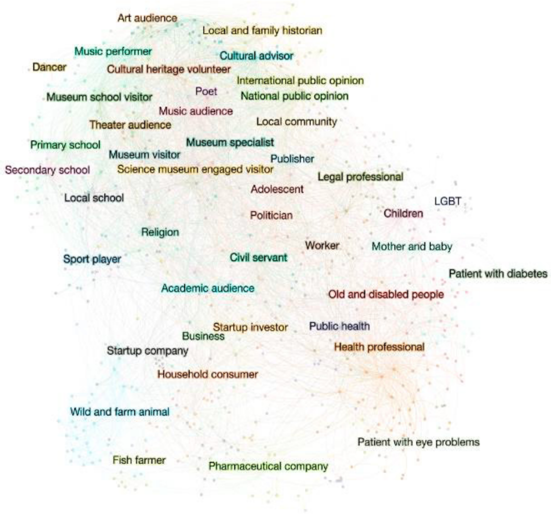


Figure 2. Map of users of research of UK universities.

Lesson learnt and future developments

In the next sub-sections, we develop some implications from this research program and discuss some extensions and new applications.

Towards the construction of indicators of research impact

The construction of indicators in S&T is a complex process, with epistemic, social and political dimensions. With respect to the epistemic dimension, there are several requirements that have been clearly outlined by the literature. Among them, we call the attention on the role of completeness.

When using statistical indicators, researchers and policy makers implicitly rely on the ability of statistical authorities to produce data that correctly refer to the universe under analysis. It is interesting to observe that in the field of S&T indicators this assumption is not warranted. This is why the overall use of data from, say, publications and patents is predicated after a standard methodological caution, of the form “we know that patents do not capture all innovation activities, but..” or “we know that indexed publications do not represent all scientific production, but…”. After a while, such clauses are omitted and often forgotten, with the exception of a stream of critical studies that remind us about the limitations of statistical indicators and of simplified input-output models.

The community of S&T indicators is reasonably cautious in accepting the use of data from Machine learning sources, such as text mining. One important reason is that it is not possible to define the reference set and establish its representativeness.

We suggest that this state of affairs can be greatly improved with the use of lexicons. Lexicons are human-made, controllable, improvable cognitive structures. If they are left open to public consultation and correction, they can incorporate corrections and updates, almost in real time. We argue that with well-designed lexicons it is possible to achieve the saturation of a semantic field. If this is true, then there are no conceptual obstacles to the construction of indicators.

If a lexicon saturates a semantic field, then the automatic extraction of whatever semantic structure in a corpus of text will deliver consistent results. They might differ depending on the specific algorithm, but the difference can be clearly explained and made transparent.

In previous research we have developed a lexicon that includes all social groups that have been hitherto considered in all social worlds in which people can be aggregated (e.g. work, profession, health, economic, social and demographic conditions, mobility, hobby, sport, entertainment, culture and education, crime), combining hundreds of published sources. We have no formal demonstration of completeness. At the same time we ask- do we have completeness, say, in patents as indicators of technology, or indexed publications as indicators of science? Certainly not, but we know reasonably well the limitations of these indicators, so that we can safely use them for a number of relevant applications.

We suggest that lexicons can do the same job for us. If we saturate the semantic field in terms of textual descriptions of relevant phenomena, then we can start to build up indicators that have proper statistical properties.

Research impact assessment in SSH

Another area in which we see clear implications from our approach is the hot debate on research impact assessment in SSH. It has been repeatedly and convincingly argued that asking SSH research to demonstrate impact in the same way as it is done for STEM is a serious mistake, with potentially far reaching implications. In particular, it might be used to justify a differential treatment of SSH and STEM in the funding of research. Under the pressure of government demand for impact and demonstrable results, decision makers and funding agencies may find it safer to maintain support for STEM research and to cut, or postpone, the support to SSH. After all, it is often said, researchers in SSH work for long time horizons and do not compete for discoveries. This means that if funds for SSH are delayed or reduced, there is no risk for the national scientific competitiveness.

With our studies we establish the following points:

- Researchers in SSH are able to identify their audiences and talk extensively (i.e. frequently) and intensively (i.e. with a variety of names) about social groups that may benefit from their research. Hence it is not true that SSH research underestimate the importance of addressing social groups as target for the impact.
- At the same time, they find more difficult to identify specifically and in a granular way their target groups, given the generality of their research.
- In order to claim that their research has produced an impact they use the same semantic structure than STEM (impact sentences with direct and indirect impact verbs + agent + mode) and the same set of verbs, implying an effort to build up causal statements.
- They however make use of a larger share of non-impact sentences in the articulation of their impact reports, given a stronger need to introduce the complex social context for the impact.
- They build longer causal chains.
- They make use of a larger number of agents.

From these quantitative findings we obtain a relatively clear picture of the differences between the ways in which research produces an impact in STEM and SSH. In order to build up a credible reconstruction of the historical pathway that has led to the impact, researchers in SSH must invoke many more agents (two times than in STEM) and describe a longer chain of causal linkages. Now from the theory of historical explanation suggested by Hammer (2008) we derive the implication that longer chains and chains with more agents are *more fragile*. This means that it is more difficult to satisfy the requirements of normativity, identifiability, manipulability and non-replaceability that are needed to establish causality in historical statements. At each juncture of the causal chain it is more difficult to credibly argue that X has indeed been a partial

sufficient condition for Y, because there might be many other factors at place or many independent agents whose actions and motivations may not well known.
Given this pattern, we advocate an approach to research impact assessment that makes full justice for these differences.

Impact mapping

Finally, we suggest a new area of application of text mining methods, i.e. mapping research institutions from the perspective of research users. This area would be complementary to the well established area of science mapping, in which institutions are represented in 2D maps on the basis of the disciplines they cultivate, or the topics that their research are addressing. We give an example below, by showing the user maps of two UK universities.

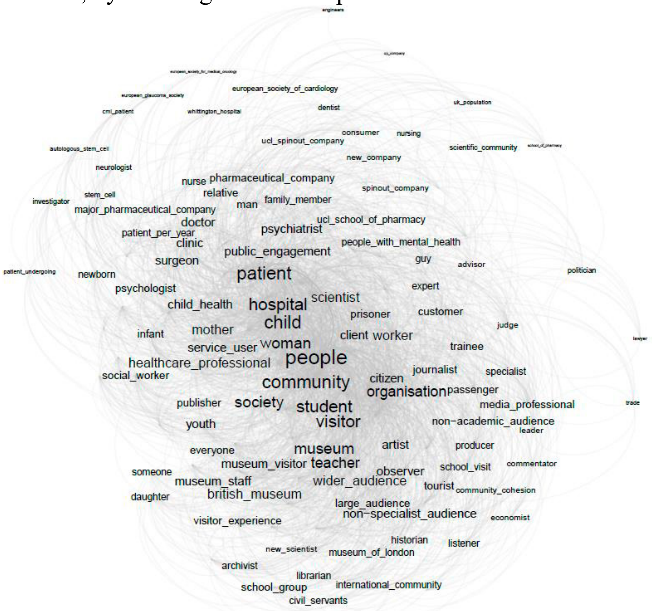


Figure 3. Map of user groups of research of University College, London.

In Figure 3 we see the map of UCL, a top research-intensive university with an international visibility. It can be observed a strong orientation towards user groups in the health sector (patient, hospital, child, woman, mother, healthcare professional community), as well an orientation towards users of cultural heritage (museum, museum visitor, student, visitor, British Museum), while on the contrary the business audience does not seem prominent.

Figure 4 shows the same map for the University of Sheffield. It seems that the main orientations here are largely different: one is towards a local audience (local school, local community, person, community, young people), another is directed towards business actors (organization, company, leader, customer, manager). Much less prominent, contrary to UCL, are the health sectors and the cultural heritage.

There will be a need to refine the analysis and perhaps to develop quantitative indicators after controlling for the robustness of classification obtains by clustering the names of user groups.

We hope there will be opportunities in the near future to extend and refine the methodologies, with an aim to improve the theoretical foundations and the methodological sophistication of research impact assessment.

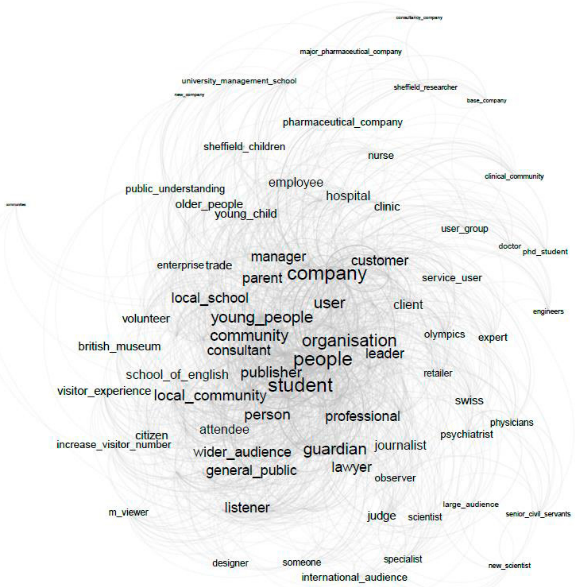


Figure 4. Map of users of research of University of Sheffield.

References

Angrist, J.D. & Pischke, J.S. (2015). *Mastering metrics. The path from cause to effect*. Princeton, Princeton University Press.

Bai, S., Zhang, F. & Torr, P. (2020). Hypergraph Convolution and Hypergraph Attention. arXiv:1901.08150.

Benneworth, P. (2015). Putting impact into context: The Janus face of the public value of Arts and Humanities research. *Arts and Humanities in Higher Education*, 14(1), 3–8.

Bertin, M., Atanassova, I., Sugimoto, C. R. & Larivière, V. (2016). The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics*, 109(3), 1417-1434.

Bonaccorsi A., Melluso, N., Chiarello, F. & Fantoni, A. (2021a). The semantic of research impact. Examining societal impact of research with Semantic Hypergraphs. Submitted for publication.

Bonaccorsi, A., Melluso, N., Chiarello, F. & Fantoni, G. (2021b). The credibility of research impact statements: A new analysis of REF with Semantic Hypergraphs. Science and Public Policy. https://doi.org/10.1093/scipol/scab008

Bonaccorsi, A., Chiarello, F. & Fantoni, G. (2021c). SSH researchers make an impact differently. Looking at public research from the perspective of users. *Research Evaluation*, 2021, 1-21, doi: 10.1093/reseval/rvab008.

Bonaccorsi, A., Chiarello, F. & Fantoni, G. (2020). Impact for whom? Mapping the users of public research with lexicon-based text mining. *Scientometrics*, https://doi.org/10.1007/s11192-020-03803-z

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey. *Journal of the Association for Information Science and Technology,* 64(2), 217-233.

Bornmann, L. & Haunschild, R. (2017). Does evaluative scientometrics lose its main focus on scientific quality by the new orientation towards societal impact? *Scientometrics*, 110(2), 937-943.

Bornmann, L. & Marx, W. (2014). How should the societal impact of research be generated and measured? A proposal for a simple and practicable approach to allow interdisciplinary comparisons. *Scientometrics*, 98(1), 211-219.

Colinet L., Joly P-B., Gaunand A., Matt M., Larédo P. & Lemarié S. (2014). ASIRPA. Analyse des Impact de la Recherche Publique Agronomique. Rapport final. Rapport préparé pour l'Inra. Paris, France.

Collier, A. (2005). Philosophy and critical realism. In G.Steinmetz (ed.) *The politics of method in the Human Sciences. Positivism and its epistemological others.* Durham, Duke University Press.

De Jong, S. P., Van Arensbergen, P., Daemen, F., Van Der Meulen, B. & van den Besselaar, P. (2011). Evaluation of research in context: an approach and two cases. *Research Evaluation*, 20(1), 61-72.

De Jong, S., Barker, K., Cox, D., Sveinsdottir, T. &van den Besselaar, P. (2014). Understanding societal impact through productive interactions: ICT research as a case. *Research Evaluation*, 23(2), 89-102.

Derrick, G. E. (2014). Intentions and strategies for evaluating the societal impact of research. Insights from REF 2014 evaluators. In: P. Wouters (Ed.), *Proceedings of the science and technology indicators conference 2014 Leiden "Context Counts: Pathways to Master Big and Little Data"* (pp. 136-144). Leider, the Netherlands: University of Leiden.

Derrick, G. E., Meijer, I., van Wijk, E. (2014). Unwrapping "impact" for evaluation: A co-word analysis of the UK REF2014 policy documents using VOSviewer. In: P. Wouters (Ed.), *Proceedings of the science and technology indicators conference 2014 Leiden "Context Counts: Pathways to Master Big and Little Data"* (pp. 145-154). Leider, the Netherlands: University of Leiden.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Digital Science (2015). *REF 2014 Impact Case studies and the BBSRC.* Available at www.bbsrc.ac.uk/documents/1507-ref-impact-case-studies-pdf/. Accessed December 3, 2019.

Digital Science (2016). *The societal and economic impacts of academic research. International perspectives on good practice and managing evidence.* Digital Research Reports, March.

Donovan, C. (2011). State of the art in assessing research impact: introduction to a special issue. *Research Evaluation,* 20(3), 175-179.

Ernø-Kjølhede, E. & Hansson, F. (2011). Measuring research performance during a changing relationship between science and society. *Research Evaluation*, 20(2), 131-143.

Gibson, A. G. & Hazelkorn, E. (2017). Arts and humanities research, redefining public benefit, and research prioritization in Ireland. *Research Evaluation*, 26(3), 199-210.

Godfrey-Smith, P. (2003). *Theory and reality. An introduction to the philosophy of science.* Chicago, University of Chicago Press.

Green, N. (2018). Towards mining scientific discourse using argumentation schemes. *Argument & Computation,* 9(2), 121-135.

Green, N.L: (2020). Recognizing rhetoric in science policy arguments. *Argument & Computation*, 1-12.

Heffernan, K. & Teufel, S. (2018). Identifying problems and solutions in scientific text. *Scientometrics*, 116, 1367-1382.

Honnibal, M. Montani, I., Van Landeghem, S. & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing. 10.5281/zenodo.1212303

Joly, P.B., Gaunand, A., Colinet, L., Larédo, P., Lemarié, S. & Matt, M. (2015). ASIRPA: A comprehensive theory-based approach to assessing the societal impacts of a research organization. *Research Evaluation,* 24, 440-453.

Kaufmann M., van Kreveld M. & Speckmann B. (2009). Subdivision Drawings of Hypergraphs. In: *Graph Drawing*, Vol. 5417, pages 396–407. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-00218-2 978-3-642-00219-9.

King's College, Digital Science (2015). *The nature, scale and beneficiaries of research impact. An initial analysis of REF (2014) impact case studies.* Research Report 2015/01. London, HEFCE.

Lawrence, J. & Reed, C. (2019). Argument mining: a survey. *Computational Linguistics*, 45(4), 765-818.

Martin, B. R. (2011). The Research Excellence Framework and the 'impact agenda': are we creating a Frankenstein monster? *Research Evaluation*, 20(3), 247-254.

Matt, M., Gaunand, A., Joly, P.B. & Colinet, L. (2017) Opening the black box of impact. Ideal-type impact pathways in a public agricultural research organization. *Research Policy*, 46, 207-218.

Menezes, T. & Roth, C. (2019). Semantic Hypergraphs. arXiv:1908.10784.

Mihalcea, R. & Tarau, P. (2004). TextRank: Bringing Order into Texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404-411.

Miettinen R., Tuunainen J. & Esko T. (2015) Epistemological, artefactual and interactional. Institutional foundations of social impact of academic research. *Minerva*, 53, 257-277.

Mikolov, T., Corrado, G., Chen, K. & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781

Molas-Gallart, J. & Tang, P. (2011). Tracing 'productive interactions' to identify social impacts: an example from the social sciences. *Research Evaluation*, 20(3), 219-226.

Morton, S. (2015). Progressing research impact assessment: A 'contributions' approach. *Research Evaluation,* 24(4), 405-419.

Olmos- Peñuela, J., Benneworth, P. & Castro-Martinez, E. (2014). Are 'STEM from Mars and SSH from Venus'? Challenging disciplonary stereotypes of research's social value. *Science and Public Policy,* 41, 384-400.

Olmos- Peñuela, J., Benneworth, P. & Castro-Martinez, E. (2015). Are sciences essential and humanities elective? Disentangling competing claims for humanities' research public value. *Arts and Humanities in Higher Education,* 14(1), 61-78.

Pearl, J. & Mackenzie, D. (2018). *The book of why. The new science of cause and effect.* New York, Basic Books.

Penfield, T., Baker, M.J., Scable, R. & Wykes, M.C. (2014) Assessment, evaluations, and definitions of research impact. A review. *Research Evaluation,* 23 (1), 21-32.

Samuel, G. N. & Derrick, G. E. (2015). Societal impact evaluation: Exploring evaluator perceptions of the characterization of impact under the REF2014. *Research Evaluation*, 24(3), 229-241.

Spaapen, J. & Van Drooge, L. (2011). Introducing 'productive interactions' in social impact assessment. *Research Evaluation,* 20(3), 211-218.

Teufel, S. & Moens, M. (2002). Summarising scientific articles --- Experiments with relevance and rhetorical status. *Computational Linguistics* 28(4), 409-446.

Woodward, J. (2003). *Making things happen. A theory of causal explanation*. Oxford, Oxford University Press.

Detecting interdisciplinarity in top-class research using topic modeling

Andrea Bonaccorsi[1], Nicola Melluso[1] and Francesco Alessandro Massucci[2]

[1] *a.bonaccorsi@gmail.com, nicolamelluso@gmail.com*
Department of Energy, Systems, Territory and Construction Engineering, University of Pisa, Pisa (Italy)

[2] *francesco.massucci@sirisacademic.com*
SIRIS Lab, Research Division of SIRIS Academic, Barcelona (Spain)

Abstract

The paper applies topic modeling to the collection of ERC-funded proposals, interim reports and relative publications, with the aim of measuring in a novel way the degree of interdisciplinarity and addressing several open research questions which broadly aim at understanding how environmental conditions can favour the blossoming of interdisciplinarity. Without venturing into potential interpretations and explanations, we present a series of quantitative results linked with the above questions, while deliberately maintaining a descriptive attitude.

Introduction

The issue of interdisciplinarity (henceforth, ID) is again on top of the table for decision makers, funding agencies, and researchers as well. And, again and again, the issue remains open and problematic. Why is ID, at the same time, so requested and appreciated by society, and so difficult to adopt, implement, recognize, and prize by funding agencies, research institutions, and researchers?

In this paper we offer two contributions to a large and growing literature. On the technical side we experiment the use of advanced text mining techniques, namely optimal topic modeling, to describe and measure ID of research (henceforth, IDR). This follows from a recognition of the literature, that emphasizes the limitations of previous efforts to measure ID with the help of metadata of publications, such as classification schemes of journals (Subject Categories) and citations. Topic modeling has been applied to ID in a few papers in recent years, with promising results. We extend this line of investigation.

On the substantive side, we aim at addressing a number of research questions that are at the core of the debate on ID and which can be summarized as follows:

a) Do we see systematic differences in the level of ID by scientific discipline?

b) Is the propensity to undertake IDR influenced by organizational factors, such as the institutional affiliation (universities vs Public Research Organizations, henceforth PROs)?

c) Is the propensity to undertake IDR dependent, in a recognizable way, on the life cycle of researchers? Are young, or mid-career, or aged researchers more inclined to submit proposals with higher levels of ID?

These questions are clearly crucial in policy making: if decision makers want to promote IDR, for example to address complex societal challenges, they must know under which conditions IDR is likely to flourish.

To tackle the above questions, we examine the research projects funded by one of the world largest funding agencies, the European Research Council (ERC). From the open archive of ERC we extract the abstract of the funded research projects in the period 2007-2018 and integrate them with interim reports that funded projects must submit during the support period as well as with the abstract of the scientific publications produced by each project and indexed either in the open repository OpenAire or in the bibliometric database Scopus. For this documentation we extract the words that describe the research project (full text of the abstract) and the list of publications cited by each scholarly work linked to each project. We build a novel dataset by reconstructing the biographic information of Principal Investigators (age, country, affiliation) and his/her list of publications (in particular, date of first publication).

We deliberately maintain an exploratory and descriptive attitude. There is large need to understand data qualitatively and with simple quantitative analysis, before entering into modeling exercises. We clearly have two limitations: due to institutional policies we can observe only winning project (so we are not in the position to run a counterfactual exercise), and submission to the ERC are not representative of the average research landscape but rather of the frontier, excellent research at the European scale.

The interest of the findings, however, will make justice of the proposed approach.

Related Works

In this section we address separately two streams of literature: the large stream of studies that try to define and measure ID (sub-sections 2.1 and 2.2) and the studies that identify possible antecedents at epistemic, institutional and individual level (sub-section 2.3).

Open issues in the definition and measurement of ID

ID has a complex meaning, that go deeply into epistemological and sociological debates. We do not enter here on these debates but follow the literature in assuming that ID is a multidimensional construct. Given the importance of the issue, there have been several efforts to identify, define and measure new indicators that might approximate the multidimensionality of the construct (Porter et al. 2006; Wagner et al. 2011). According to a classical conceptualization diversity can be defined in terms of (i) variety (number of disciplines); (ii) balance (distribution of disciplinary contributions); and (iii) disparity (distance or degree of difference between contributing disciplines) (Rao, 1982; Stirling, 2007). These definitions may lead to remarkably different measures of ID (Wang et al. 2015).

A consistent literature has defined ID by applying measures of diversity to the distribution of articles and citations by Subject Category (Morillo et al. 2001). According to this approach, the journals cited in the reference list of articles are classified by Subject Category and a citing-cited matrix is generated. After defining the metric of diversity it is possible to reduce the dimensionality of the resulting matrix and to project the disciplinary profiles on a map (Rafols et al. 2010; Carley and Porter, 2012; Cassi et al. 2017; Carley et al. 2017).

Following another approach, the diversity of disciplines is defined by the co-citation of articles in journals in different Subject Categories (Moya-Anegón et al. 2004; 2007; Porter et al. 2019). The most common definitions of diversity are the Rao-Stirling measures (Rao, 1982; Stirling, 1997), the Zhang et al. true diversity (Zhang et al. 2016), in addition to the standard measures of cosine distance (Porter et al. 2007; 2008) and Jaccard distance. Different measures of diversity place different emphasis to its dimensions and have different statistical properties, which are the object of critical examination. Klavans and Boyack (2009) have compared systematically the diversity measures and the resulting maps and support the use of cosine distance (see also Leydesdorff and Rafols, 2012). Ciotti et al. (2016) criticize the use of Jaccard measure as it is not independent on the size of the discipline. Fontana et al. (2018) support this view (i.e. the probability of integration between disciplines depends on the absolute and relative size of disciplines) and add a dynamic dimension (i.e. the probability of integration depends on the growth of disciplines). These contributions make it clear that the adoption of alternative definitions of ID and of diversity measures may lead to very different outcomes. As Leydesdorff and Rafols (2011) put it, “different indicators may capture different understandings of multifaceted concepts”.

The Topic Modeling approach to ID

Partially in response to the limitations of the literature based on the structure of citations a recent stream of literature has been exploring the potential of Text mining techniques, applied to the words appearing in titles, abstract, or keywords of articles, less frequently on the full text. By using a variety of algorithms (in particular, Latent Dirichlet Allocation, LDA: Blei et al. 2002) and clustering procedures (Topic modeling) it is possible to represent the content of scientific papers (Rosen-Zvi et al., 2004; Lu and Wolfram, 2012; Gerrish and Blei, 2010; Wang et al. 2011; Nichols, 2014). In this case there is no need for ex ante classification, as in the Subject Category case, so that all limitations of the classification are removed. ID can be defined directly on the maps of topics.

The idea behind topic modeling as a technique for measuring ID has been taken on board by the National Science Foundation (henceforth, NSF) in the USA. According to NSF, the advantage of topic model approach on measuring ID is that its algorithm uses data drawn from the research itself, rather than from institutional structures through which the research was produced (Nichols, 2014).

Lu and Wolfram (2012) compare citation-based and word-based methods with topic-based techniques. By using LDA it is possible to uncover relationships between topics that would remain otherwise hidden due to terminological classifications. Also peripheral authors with few citations might be found central in terms of new topics. At the same time they call attention to the limits of topic modeling: there is no optimal rule for the determination of the number of topics, the label of the topic is assigned by convenience by the researcher, and the units of observation are single words and not more complex and long semantic units.

We emphasize that the methodological debate is not just of technical interest, but is crucial also for the hot debate on ID that takes place at epistemological level, as well as in research assessment and policy making. If measures of IDR are flawed, they deliver wrong messages to decision makers. There is therefore a keen interest in contributing to the methodology.

Why ID so difficult?

As stated above, ID is paradoxical, in the sense that there is large gap, so to say, between its demand, which is sustained and growing, and its supply, which is scarce. In this sub-section we explore three main areas that have been the object of disparate contributions, in which there are potential antecedents for this gap. They are: (a) epistemic, or disciplinary effects; (b) institutional effects; (c) individual, career-related effects.

Epistemic and disciplinary effects

Huutonieni et al. (2010) make the important distinction between addressing the issue from an instrumental and pragmatic perspective or from a more long term, conceptual, cognitive and epistemological perspective (Boon and van Baalen, 2019). From the latter perspective it must be admitted that we do not have a full scale theory of the intensity in which ID is practiced in different disciplines and of the underlying reasons. Several authors offered important insights, but no general results (Silva et al., 2013).

Somewhat less is known about ID in Social Sciences and Humanities. Garnier et al. (2013; 2014) evaluate a NSF program (Human and Social Dynamics) explicitly aimed at promoting ID within SSH and STEM (labeled cross-disciplinarity). They find a change in an integration measure before and after the program (2004-2008) and argue that without a specific program it would not have taken place spontaneously. They therefore recommend the adoption of targeted funding and specific requirements in grant calls.

On the other hand, researchers in SSH tend to cite contributions from STEM in an increasing way. Liu et al (2017) document a sharp increase in STEM citations in SSH papers in China but warn that there is an inverse U-shaped relation between the extent of interdisciplinary citations and the scientific impact.

Summing up, we expect large differences across disciplines, but we do not have a theory supporting specific predictions.

Institutional effects

The literature has raised a number of concerns on the relation between institutional features of science systems and ID. Let us distinguish between funding institutions (government, funding agencies) and performing institutions (universities, PROs). With respect to funding, several authors have shown that IDR has consistently lower funding success (Metzger and Zare, 1999; Bromlsam et al. 2016; Kwon et al. 2017). Various explanations can be offered: panel of reviewers have less experience, the matching between projects and assessor expertise is more difficult, and it is easier to justify and explain the funding of conventional projects. Metzger and Zare (1999) strongly suggested the adoption of targeted funding programs.

With respect to performing institutions, much less is known. Universities may be at advantage since they have large variety of disciplines. In principle, social interaction among colleagues of different background may be facilitated (Cummings and Kiesler, 2005). Some studies have compared universities with a broad coverage (generalist universities) with dedicated or specialist universities (mainly in applied fields). Bonaccorsi and Secondi (2017) find that generalist universities outperform specialist ones in research productivity across all European countries (Bonaccorsi et al. 2021a; 2021b). In any case, these studies do not measure ID but overall productivity. On the other hand, when IDR requires prolonged interaction over time, PROs may be more flexible in arranging dedicated structures, relaxing researchers from the constraint of teaching. Also in this case we must explore the issue without a strong guidance from the literature. We are able to examine whether researchers who submit more interdisciplinary projects are affiliated to universities or PROs.

Age and career effects

An intriguing issue is the relation between the propensity of scientists to engage into IDR (either individually or within interdisciplinary teams) and their research career. We can distinguish two intertwined issues: a cognitive or epistemic issue, and an institutional or incentive issue. In some sense, they are the combination, at individual level, of the two powerful forces examined above, namely epistemic and institutional effects.

At the cognitive or epistemic issue we ask under what conditions do individual researchers conceive research ideas that require interdisciplinary knowledge to be addressed scientifically. Do they have deep issues in mind since their early years and develop research programs to address them later during their career? Or rather, do they focus initially on a narrow disciplinary focus and learn only later the need for opening their mind to other disciplines?

The cognitive or epistemic issue is inextricably linked to the problem of incentives. Is IDR good or evil for early stage researchers? Is the academic career enhanced or damaged by IDR?

Here the prevailing literature has been fairly sharp in proposing that the institutional systems based on publish or perish places severe penalties to IDR (Rafols et al. 2012; Chen et al. 2015; Wang et al. 2017). Other authors offer a more articulated view. For example, Lariviére and Gingras (2010) show that there is not a general pattern of correlation between ID and citations. There are large differences across disciplines. More recently, Okamura (2019) has challenged the negative view by showing that, in the population of highly cited papers (top 1% annual citation counts) increasing by one the number of disciplines leads to 20% increase in field normalized citations.

This debate is also important for policy makers: if they want more IDR but IDR is penalized in the academic career, simple financial schemes may not be enough.

Data and Methodology

Data

The analysis is conducted on the abstracts of the ERC scientific production. In particular, we focused on analysing the projects funded by the program from 2007 to 2018 and the publications which acknowledge financial support by those projects. This textual corpus is cross-linked with information on the respective researchers (i.e., to the PIs of the different projects) and with information about the institutions where those projects were effectively carried out. Since the funding scheme of ERC is conceived within an open access framework, we could make use of a series of open platforms to retrieve the above information: UNiCS[1], CORDIS[2], OpenAire[3] and ETER[4] (Mosca et al., 2018). These platforms give us the possibility to gather data about (i) the projects funded, (ii) the publications published thanks to the funding project, (iii) the principal investigators of the projects and (iv) the hosting institution of the project. Moreover, we gathered additional data at the level of publication, researcher and institution. With respect to publications, we extracted: the average reference age (namely the average distance between the publication year of a document and the publication year of the references) and the subject areas assigned by Scopus. For what concerns researchers, we collected their academic biographies (using Scopus) and we measured the distance (in year) from the winning year of the project to the year of their first publication (as PhD). Finally, for what concerns institutions, we gathered the foundation year (using ETER).

From topic models to ID

The measure of ID relies on the definition of an ID index calculated from the results of the topic modelling applied on the ERC scientific corpus. In particular, after removing the stop-words and stemming the texts, we used the LDA method to extract the topic models. As well as other unsupervised algorithms, LDA required a careful assessment of the reliability of the model resulted from the computation. In this case, we performed a concurrent analysis of 4 different metrics (Griffiths et al. 2004, Cao et al. 2009, Arun et al. 2010, Deveaud et al. 2014) to choose the best number of topics.

In LDA the topic distribution is assumed to have a sparse Dirichlet prior distribution (Ng et el., 2011). Considering the goal of the present paper, given a set of D documents and a chosen number of T topics, the main output of the LDA is a matrix of probability distribution (β_{dt}) that quantifies the probability β of a document d of belonging to the topic t. In this framing, an interdisciplinary topic would be one that groups documents spanning several pre-defined disciplinary classifications: for this reason, we shall consider the panel structure of the ERC program (which is structured into 3 main areas, namely Life Sciences, Physical Sciences and Engineering and Social Sciences and Humanities, which branch in turn into 25 panels) as the main reference point to calculate an index that explains the ID of a topic. When a principal investigator submits a project, she or he chooses a panel to which the contribution of the research belongs.

So that, given the Dirichlet distribution resulted from the topic models, we follow the assumptions of Leydesorff (2011): we determine the diversity of topics measuring how they are distributed among the three panels. Thus, the ID Index of a topic S_t is calculated as follows:

$$S_t = \frac{H(t)}{\log\ 3} = \frac{-(\gamma_{PE,t}\log(\gamma_{PE,t}) + \gamma_{SH,t}\log(\gamma_{SH,t}) + \ \gamma_{LS,t}\log(\gamma_{LS,t}))}{\log 3}$$

[1] https://unics.cloud/

[2] https://cordis.europa.eu/

[3] https://www.openaire.eu/

[4] https://www.eter-project.com/#/home

H(t) is the Shannon Entropy calculated among the probabilities that topic *t* belongs to the panels. PE, SH and LS represent respectively the "Physics and Engineering", "Social and Humanities" and "Life Sciences" panels.

Given the ID of a topic, it is possible to employ the document-topic probability matrix β_{dt} to calculate the ID of documents, researchers, institutions and disciplinary fields as follows:

$$S_d = \sum_{t=1}^{T} S_t \beta_{dt}; \qquad S_r = \frac{1}{n_r}\sum_{d=1}^{n_r} S_d; \qquad S_i = \frac{1}{n_i}\sum_{d=1}^{n_i} S_d; \qquad S_s = \frac{1}{n_s}\sum_{d=1}^{n_s} S_d;$$

S_r measures the ID a document *d*. S_r measures the ID of a researcher, with n_r equals the total number of documents produced by the researcher *r*. S_i measures the ID of an institution, with n_i the total number of documents produced by the institution *i*. S_s measures the ID of disciplinary field, with n_s the number of subject areas that Scopus attributes to the document. Note that S_r and S_i are calculated considering only the abstracts of the projects, while S_s is calculated considering only the publications.

Main findings

Descriptive statistics

We gathered a total sample of 41,379 abstracts (9,365 projects and 32,014 publications) and data about 1,062 institutions and 7,056 researchers. After performing the LDA algorithm, we evaluate the results identifying the best number of topics according to the assessment of state-of-the-art metrics (see crf. "*From topic models to ID*"). The best model resulted in 200 topics. The projects funded by the ERC and their related publications have an average degree of ID at around 0.45. The average degree grows over time. Figure 1 shows that the average degree is between 0.43 and 0.44 in 2008-2010 and 0.46 in 2018. We have no comparison with other institutions and/or periods, given the novelty of the indicator adopted. Nevertheless, the data show a slight increase in the decade if one discards the fluctuations of 2015-2016, which are generated by the change in Framework program between FP7 and H2020.

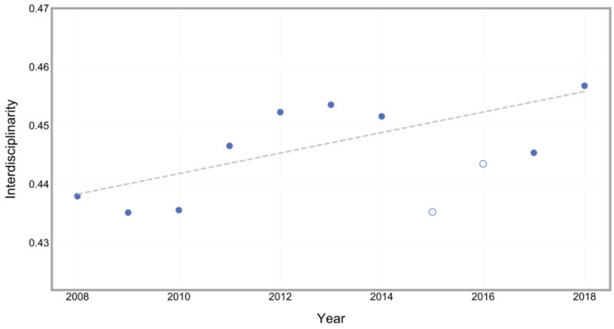


Figure 1. Dynamics of the ID degree. Year 2008-2018.

Disciplinary effect

There are large differences across disciplines in the degree of ID. The largest values are found in Physical and Engineering sciences (PE) in PE6 and PE10. On average, the degree of ID of PE fields is comparable to Life Sciences (LS).

Interestingly the lowest values are found in Social Sciences and Humanities (SH). The least interdisciplinary projects are found in SH2, SH5 and SH6, with SH1 at short distance.

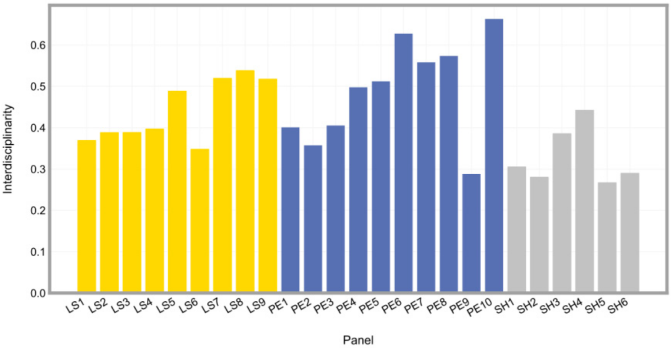


Figure 2. Degree of ID across disciplinary areas in the ERC panel.

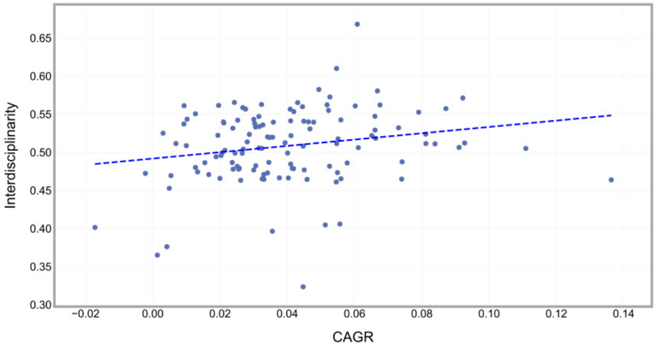


Figure 3. Relation between the rate of growth of publications (Compounded Annual Growth Rate 2008-2018) in the main field of the research project and the degree of ID.

There is a positive, although moderate, relation between the rate of growth of scientific disciplines and the degree of ID (Figure 3).

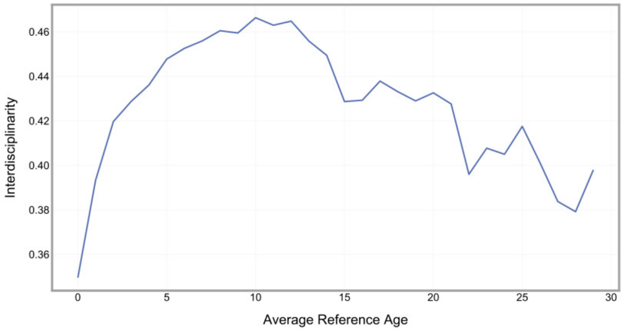


Figure 4. Relation between the average age of the papers cited in the ERC projects and the degree of ID.

We can interpret this finding in various ways. Fast growing disciplines may offer more opportunities for exploring the intersections and boundaries with other disciplines. Or, rather, they may result from the merge or convergence between previous established disciplines (as in the classical case of nanoscience), so that they maintain a dialogue with the originating fields. Or there is a demographic factor at place: fast growing fields may attract young researchers who are more oriented towards ID. For the time being, we cannot disentangle these effects.

On the contrary, we have a sharp result with respect to the structure of citations of other papers that the winners of ERC projects place in their documents. The list of publications cited in the abstract of the project can be considered a knowledge tree that makes explicit the origins of the research idea, as well as the state of the art that the applicants aim at innovating. What we find is that the projects with the largest degree of ID cite papers that are, on average, 10-11 year old (Figure 4). In other words, more interdisciplinary projects do not cite very old papers, but also do not cite very young papers. The former finding is perhaps trivial: applicants to ERC aim at improving over the state of the art and it would be difficult to believe that it includes very old pieces of knowledge. If they take as reference very old state of the art, it means that the field is not very dynamic (perhaps with the important, but very small, exception of unsolved mathematical problems, in the venerated tradition of Hilbert's list).

The latter finding is, on the contrary, surprising and interesting. Successful researchers that operate at the frontier do not use in their proposals and in the interim reports the very last papers. Is this in contradiction with the notion of frontier research? We offer an epistemic interpretation, as follows. ERC applicants do know they will be evaluated by top scientists in all disciplines and know that very young results need to be validated. The initial validation is likely to be done within disciplinary boundaries. There seems to be a *gestation period*, in which young results are progressively validated by their own discipline and become known outside the discipline. Remember that we compute the average age of the cited papers. The list of references may well include very young papers, but they are embedded into a knowledge structure which is relatively mature.

Institutional effects

It is perhaps true that PROs claim they support more IDR than traditional universities. The plain truth is that there is almost no difference between the two types of affiliation (Figure 5). No statistical test needed. [5]

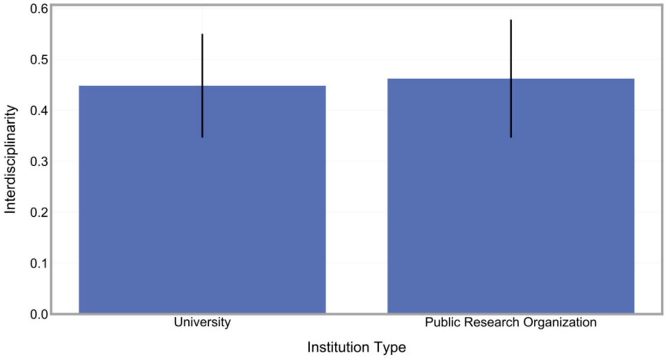


Figure 5. Average degree of ID by type of affiliation of ERC applicant.

In the absence of a robust theory that establishes a linkage between the institutional type (university vs PRO) and the propensity to engage into IDR, we leave this result to open discussion. On the contrary, we find moderate evidence that younger institutions are more likely to be the affiliation of ERC applicants with a higher degree of ID. With few exceptions, a high average degree of ID, as well as a larger variance across projects, is found in institutions established in the last two centuries (Figure 6). Combining the findings from Figure 5 and 6 it seems that the nature of the institution does not matter, but the age of the institution plays some

[5] Anyway, the ANOVA test of difference of means is rejected (p-value 0.278197).

role. We might be tempted to conclude that younger institutions try to differentiate from established institutions by adopting and nurturing a more interdisciplinary approach. The evidence however is only suggestive: conclusions are premature.

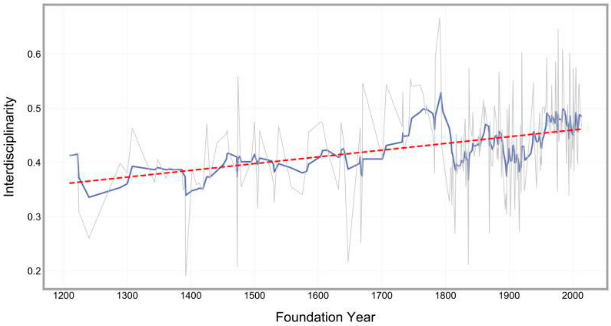


Figure 6. Relation between the age of the affiliation (foundation year of the institution) of ERC applicants and degree of ID.

Individual effects

As stated above, we have collected the biographies of the ERC winners and created a variable that measures the distance (in year) from the year of the PhD. This is an important information, since it dictates formally the eligibility of candidates to different ERC schemes. In particular, Starting Grants are dedicated to young researchers whose PhD degree has not been awarded more than 7 years before, while Advanced Grants do not place any constraint on the age of Principal Investigator. As it can be seen in Figure 8 the distribution of age of applicants has a modal value at 17 years after the PhD, while the bulk of the distribution is in the 10-17 years after the PhD. Researchers applying for Advanced Grants are not young post-doc, but rather young established scholars.

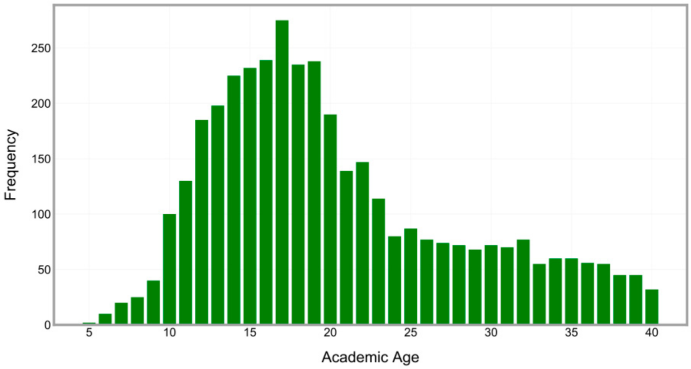


Figure 7. Distribution of ERC winners by academic age (number of years after the PhD).

Having such a large range in the distribution of academic age of ERC winners it is interesting to investigate whether more interdisciplinary projects are proposed by younger or older researchers. What we find is again striking: the peak of ID is achieved in projects in which the PI is 25 year older than the PhD, that is, is found in his/her ‘50s.

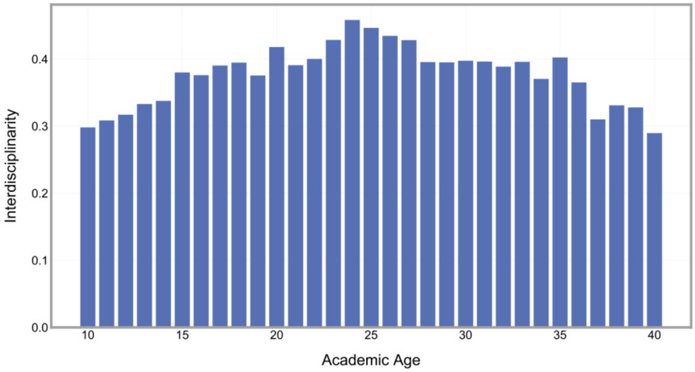


Figure 8. Distribution of the academic age of ERC applicants (number of years since PhD) by degree of ID.

Discussion and conclusions

The notion that IDR arises from young research and from young researchers is not supported. Young researchers must demonstrate first their contribution to their own discipline, focusing on internal topics, and react to the career and incentive systems. It seems that successful scientists that engage into highly interdisciplinary projects have already established their academic excellence. Whether this excellence has been gained with purely or interdisciplinary projects is not observable with our data and must be kept open. However, if ID would be an attribute of researchers, or a personal attitude, then we would see it at any age in the selected proposal. On the contrary, the degree of ID grows very little in the interval between 10 and 15 year of academic age.

An alternative conceptualization, which we find more persuasive, is a sort of life-cycle theory of individual propensity to ID. This conceptualization combines insights on the age of researchers and the age of cited literature. It goes as follows. All researchers are born disciplinary. Some of them maybe cultivate in their mind very audacious research questions that require out-of-the-box research approaches. But they do not take the risk of making them the core of their scientific production at the initial stages of the career. They know they are evaluated by senior colleagues on the basis of their contribution to the discipline. If they are successful they gain reputation and recognition. By going deeper and deeper into scientific problems they may come to appreciate contributions from other disciplines. They devote more time to read journals that are rarely cited by their colleagues but promising. They engage more freely into discussions without disciplinary boundaries. At some point in time they understand that entering into other fields, or collaborating actively with researchers from other fields, is promising, intellectually rewarding, creative, and original. They understand that the disciplinary recognition of this work will be delayed, because most of their colleagues will not accept or appreciate the break of boundaries. Nevertheless, they are no longer under the threat of delivering contributions that are immediately recognized by colleagues. They are established. Maybe their reputation may be at stake, not their tenure or salary. An intriguing question is how to verify whether researchers with a strong interdisciplinary orientation had this approach since their early years- but they did not show it in projects and publications. This would require additional research effort.

References

Arun, R., Suresh, V., Madhavan, C. V. & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observation. In: *Pacific-Asia conference on knowledge discovery and data mining,* (pp. 391–402). Springer, Berlin, Heidelberg.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research,* 3, No. Jan, pp. 993-1022.

Bonaccorsi, A. & Secondi, L. (2017). The determinants of research performance in European universities: a large scale multilevel analysis. *Scientometrics*, 112(3), 1147-1178.

Bonaccorsi, A., Belingheri, P. & Secondi, L. (2021a). The research productivity of universities. A multilevel and multidisciplinary analysis on European institutions. *Journal of Informetrics*, 15, 101129.

Bonaccorsi, A., Belingheri, P. & Secondi, L. (2021b). Economies of scope between research and teaching in European universities. Submitted for publication.

Boon, M., & van Baalen, S. (2019). Epistemology for IDR. Shifting philosophical paradigms of science. *European Journal for Philosophy of Science*, 9:16.

Bromlsam, L., Dinnage, R. & Hua, X. (2016). IDR has consistently lower funding success. *Nature*, 534, 684-687.

Carley, S. & Porter, A.L. (2012). A forward diversity index. *Scientometrics*, 90(2), 407–427.

Carley, S., Porter, A.L., Rafols, I. & Leydesdorff, L. (2017). Visualization of disciplinary profiles. Enhanced science overlay maps. *Journal of Data and Information Science*, 2(3), 68-111.

Cassi, L., Champelmont, R., Mescheba, W., & de Turckheim, E. (2017). Analysing institution ID by extensive use of Rao-Stirling diversity index. *PLoS ONE*, 12(1), e0170296.

Cao, J., Xia, T., Li, J., Zhang, Y. & Tang, S. (2009). A density-based method for adaptive LDA model selection", Neurocomputing. 72(7–9), 1775–1781.

Chen, S., Arsenault, C. & Lariviére, V. (2015). Are top-cited papers more interdisciplinary? *Journal of Informetrics*, 9(4), 1034-1046.

Ciotti, V., Bonaventura, M., Nicosia, V., Panzarasa, P. & Latora, V. (2016). Homophily and missing links in citation networks. *EPJ Data Science*, 5:7.

Cummings, J.N.& Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science*, 35(5), 703-722.

Deveaud, R., SanJuan, E. & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique,* 17(1), 61–84.

Fontana, M., Iori, M., Montobbio, F. & Sinatra (2018). A bridge over troubled water. Interdisciplinarity, novelty and impact. Università Cattolica del Sacro Cuore, Dipartimento di Politica Economica, *working paper* 2018/2.

Garnier, J., Porter, A.L., Borrego, M., Trau, E. & Tentonico, R. (2013). Facilitating social and natural science cross-disciplinarity. Assessing the human and social dynamics program. *Research Evaluation*, 22(2), 134-144.

Garnier, J., Porter, A.L. & Newman, N.C. (2014). Distance and velocity measures. Using citations to determine breadth and speed of research impact. *Scientometrics*, 100, 687-703.

Gerrish, S.M. & Blei, D.M. (2010). A language-based approach to measuring scholarly impact. ICML'10: *Proceedings of the 27th International Conference on Machine Learning*. June, 375-382.

Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(1), 5228–5235.

Huutonieni, K., Klein, J.T., Bruun, H. & Hukkinen, J. (2010) Analyzing ID. Typology and indicators. *Research Policy*, 39, 79-88.

Klavans, R. & Boyack, K.W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, 60(3), 455-476.

Kwon, S., Solomon, G.E.A., Youtie, J. & Porter, A.L. (2017) A measure of knowledge flow between specific fields. Implications of ID for impact and funding. *PLoS ONE*, 12(10), e0185583.

Lariviére, V. & Gingras, Y. (2010). On the relationship between ID and scientific impact. *Journal of the American Society for Information Science and Technology*, 61(1), 126-131.

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the ID of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9), 1303-1309.

Leydesdorff, L. & Rafols, I. (2011). Indicators of the ID of journals. Diversity, centrality, and citations. *Journal of Informetrics*, 5, 87-100.

Leydesorff, L. & Rafols, I. (2012). Interactive overlays. A new method for generating global journal maps from Web of Science data. *Journal of Informetrics*, 6, 318-332.

Liu, M., Shi, D. & Li, J. (2017). Double-edged sword of interdisciplinary knowledge flow from hard sciences to humanities and social sciences. Evidence from China. *PLoS ONE*, 12(9), e0184977.

Lu, K. & Wolfram, D. (2012). Measuring author research relatedness. A comparison of word-based, topic-based and author cocitation approaches. *Journal of the American Society for Information Science and Technology*, 63(10), 1973-1986.

Metzger, N. & Zare, R.N. (1999). IDR. From belief to reality. *Science*, 283(5402), 642-643.

Morillo, F., Bordons, M. & Gómez, I. (2001). An approach to ID through bibliometric indicators. *Scientometrics*, 51(1), 203-222.

Moya-Anegón, F., Vargas-Quesada, B., Chinchilla-Rodriguez, Z., Corera-Álvarez, E., Munoz-Fernández, F.J. & Herrero-Solana, V. (2007). Visualizing the marrow of science. *Journal of the American Society for Information Science and Technology*, 58(14), 2167-2179.

Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodriguez, Z., Corera-Álvarez, E. & Munoz-Fernández, F.J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, 61(1), 129-145.

Mosca, A., Roda, F. & Rull, G. (2018). UNiCS-The Ontology for Research and Innovation Policy Making. *Frontiers in Artificial Intelligence and Applications*, 306, 200-207.

Nichols, L.G. (2014). A topic model approach to measuring ID at the National Science Foundation. *Scientometrics*, 100, 741-754.

Porter, A.L., Carley, S., Cassidy, C.N., Youtie, J., Schoeneck, D.J., Kwon, S. & Salomon, G.E.A. (2019). Measuring IDR categories and knowledge transfer. A case study of connections between cognitive science and education. *Perspectives on Science*, 27(4), 582-618.

Porter, A.L., Cohen, A.S., Roessner, J.D. & Perreault, M. (2007) Measuring researcher ID. *Scientometrics*, 72(1), 117-147.

Porter, A.L. & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, 81(3), 719-745.

Porter, A.L., Roessner, J.D., Cohen, A.S. & Perreault, M. (2006). IDR: Meaning, metrics and nurture. *Research Evaluation*, 15(3), 187-195.

Porter, A.L., Roessner, J.D. & Heberger, A.E. (2008). How interdisciplinary is a given body of research? *Research Evaluation*, 17(4), 273-282.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P. & Stirling, A. (2012). How journal rankings can suppress IDR. A comparison between innovation studies and Business & Management. *Research Policy*, 41, 1262-1282.

Rafols, I., Porter, A.L.& Leydesdorff, L. (2010). Science overlay maps. A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, 61(9), 1871-1887.

Rao, C.R. (1982). Diversity. Its measurement, decomposition, apportionement and analysis. *Sankhya: The Indian Journal of Statistics*, 44(1), 1-22.

Rosen-Zvi, M., Griffiths, T., Steyvers, M. & Smith, P. (2004). The Author-topic model for authors and documents. *Proceedings of the UAAI Conference*, 487–494.

Silva, F.N., Rodriguez, F.A., Oliveira, O.N. & da Costa, L. (2013). Quantifying the ID of scientific journals and fields. *Journal of Informetrics*, 7, 469-477.

Stirling, A. (1997). On the economics and the analysis of diversity. *SPRU Working Paper*, no. 28.

Wagner, C., Roessner, J., Bobb, K., Klein, J., Boyack, K., Keyton, J., et al. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, 5(1), 14–26.

Wang, H., Ding, Y., Tang, J., Dong, X., He, B., Qiu, J., et al. (2011). Finding complex biological relationships in recent PubMed articles using Bio-LDA. *PLoS One*, 6(3), 1–14.

Wang, J., Thijs, B. & Glänzel, W. (2015). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PLoS ONE*, 10(5): e0127298

Wang, J., Veugelers, R. & Stephan, P. (2017). Bias against novelty in science. A cautionary tale for users of bibliometric indicators. *Research Policy*, 46, 1416-1436.

Zhang, L., Rousseau, R. & Glänzel, W. (2016). Diversity of references as an indicator for ID of journals. Taking similarity between subject fields into account. *Journal of the American Society for Information Science and Technology,* 67(5), 1257-1265.

Stability through Constant Turnover: The Replacement Rate of the Scientific Workforce

Clara Boothby [1], Staša Milojević[2], Vincent Larivière[3], John Walsh[4], Filippo Radicchi[5] and Cassidy Sugimoto[6]

[1] *crboothb@indiana.edu*
Indiana University, Department of Informatics, Bloomington, IN (USA)

[2]*smilojev@indiana.edu*
Indiana University, Department of Informatics, Bloomington, IN (USA)

[3]*vincent.lariviere@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, Montréal, Québec (Canada)

[4]*john.walsh@pubpolicy.gatech.edu*
Georgia Tech, School of Public Policy, Atlanta, GA (USA)

[5]*filiradi@indiana.edu*
Indiana University, Department of Informatics, Bloomington, IN (USA)

[6]*sugimoto@indiana.edu*
Indiana University, Department of Informatics, Bloomington, IN (USA)

Abstract

Training new scientists is essential to ensure that the scientific workforce is sustained and able to produce new findings in the future, but academic careers are shortening for each successive cohort. We aim to connect the shortening of academic careers to larger trends in the growth and turnover rates of the academic workforce. Using publishing data to estimate the distribution of career ages in the academic workforce of the United States, we find that the number of researchers entering the workforce is consistently scaled over time to replace leaving researchers at a ratio of 1.27, indicating a continued exponential growth of the workforce. However, as a large proportion of researchers depart academic science in the early career stages, we argue that the population of the early career workforce turns over every few years to fulfill essential positions within academic research rather than to replace retiring researchers. The implications of this may be that new researchers are trained to fulfil necessary research roles vacated by departing early career scientists rather than to ultimately achieve long-term tenure track positions.

Introduction

Training new scientists is essential to ensuring that the scientific workforce is sustained and able to make advancements. For established researchers, the mentorship of young researchers allows for the reproduction of the scientific community, and the doctoral degree is often considered the mechanism through which established researchers produces the next generation that will replace them as they retire (Larson et al., 2014; Malmgren et al., 2010). As PhD trained scientists face an overwhelmingly competitive job market for long-term academic positions, many observers argue that these issues may be the result of over-training based on faulty assumptions about the number of academic positions available for trained PhDs (Bourne, 2013; Perlstein, 2014). However, it is possible that these conditions are endemic to the structure of the academic workforce, particularly if early career researchers are an essential working population in it. As the pressure for scientists to publish frequently mounts, adding qualified team members to scientific endeavors may also be seen as a way to increase productivity.

When considering the context of team science, the fact that researchers in training frequently fulfill crucial roles on the research team may be a greater immediate motivation for scientific training, rather than the speculation that new faculty will eventually be needed to replace retiring professors. Previous studies indicate that graduate students have contributed to over

30% of papers published in Canada, indicating that they contribute considerably to the advancement of science during their training (Larivière, 2012). The division of labor in scientific working groups also often delegates essential research tasks, including lab maintenance, data collection, and analysis, to PhD students and postdoctoral researchers (Holden, 2015; Larivière et al., 2016; Robinson-Garcia et al., 2020; Walsh & Lee, 2015). In this way, the full participation of early career scientists in publication suggests that research training itself is entwined with the motivations of team-based science. Here, we use academic publishing data, we examine the dynamics of entrance and exit from the academic wing of the scientific workforce of the United States, and we connect concerns about the shortening of academic careers to trends in the growth and turnover rates of the academic workforce.

Methods

Publication data for this paper was obtained publication from the Web of Science (WOS), to which the author disambiguation algorithm from the Center for Science and Technology Studies (CWTS) at Leiden University was applied (Caron & Van Eck, 2014). Data covers the years 1980 to 2018. As academic systems vary dramatically across countries, we focused only on the US scientific workforce, which is defined as researchers for whom at least 85% of their papers have a US affiliation. The entire publication history of these researchers (N=3,525,183) was considered, including papers published outside the US.

For each year, we identify the researchers for whom this is the last year that they appear on any publications as a proxy for the end of their academic research career. To adjust for disambiguation errors, we filtered out authors who had only published one paper. We inferred the primary field for each researcher based on the broad subject category for the majority of each authors' papers; in the case of ties, the first field to appear is assigned as the primary field given our particular focus on early careers. Because that the dataset starts in 1980, the number of researchers identified as starting their careers in the years shortly after 1980 are over-represented. Similarly, after 2010 an increasing proportion of active authors would be misclassified as leaving academic science when their last publication is a couple years before the end of the dataset. To account for this, we limit the analysis to 1990-2010.

Results

Consistent growth in science and the replacement rate

We find that for each year between 1990 and 2010, the ratio of publishing researchers entering the overall academic workforce (publishing their first paper) to those leaving (publishing their last paper) stays consistently around a mean value of 1.27 (Fig. 1). We define this ratio as the replacement rate, meaning that in aggregate for every person leaving the academic workforce, an average of 1.27 people enter it. This finding remains consistent with the established finding that the size of the academic workforce increases at an exponential rate, as the ratio is above the equilibrium point of 1.

Different fields exhibit different behavior, and Figure 1 also shows the replacement rate for a selection of fields: Physics (1.16), Health (1.60), and Engineering and Technology (1.25). Physics demonstrates a gradual decrease in its ratio over time, indicating a slowing of growth between 1990 and 1998. Physics' replacement rate stands in contrast to Health, which exhibits a much higher stable ratio than any other field. Engineering and Technology exhibits a growth rate that is consistent with the overall growth rate of all fields in the dataset, but with greater annual fluctuation. The consistency of the overall ratio of entering researchers to leaving researchers suggests that the science workforce exhibits a stable capacity for growth, such that the entry of researchers in training is balanced by a proportionate number of researchers leaving academic environments that require regular publication.

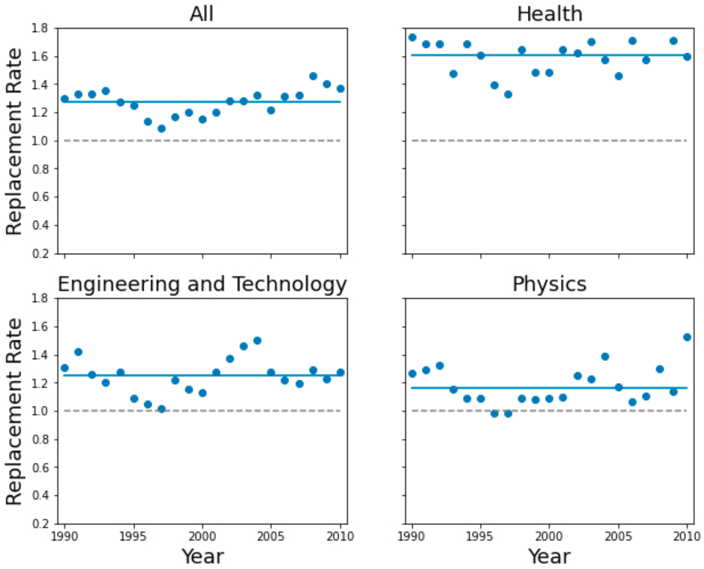


Figure 1. The ratio of researchers entering to leaving the workforce in each year, both overall and in a selection of 3 fields. The mean value (solid line) and the equilibrium value at 1.0 (dashed line) are depicted for comparison.

Turnover rates in academic science

The turnover rates of scientific fields indicate the proportion of each field in academic science who will be exiting academic science in each year, whose position will need to be replaced in the following year. Even as the size of the academic workforce grows exponentially, the mean turnover rates for all fields (0.15) and Engineering and Technology (0.16) both remain stable over time, with standard deviations of 0.008 and 0.012 respectively and without significant increasing or decreasing trends (Fig. 2A).

While we have highlighted Engineering and Technology as an example over time, we also find differing turnover rates for the extended field list (Fig. 2B), which are suggestive of field differences in team structure. Chemistry (0.18), Engineering and Technology, and Biomedical Research (0.17) exhibit the three highest field turnover rates, which may indicate a faster cycling of researchers through positions in the academic research team. The lowest turnover rates in Mathematics (0.10), Social Science (0.10), and Earth and Space (0.11) may indicate that research positions in these fields are more stable or that researchers tend to be in longer term positions when they begin authoring academic papers.

Share of researchers departing academic science from each career age

As the entry of new researchers is balanced against the departure of existing researchers, we investigated the career ages at which researchers stop publishing, presumably leaving the academic workforce. Out of the population of scientists departing academic research in each year, the share of departing researchers from each career age increases as the career age decreases (Fig. 3).

In all presented years, researchers with career ages of 2 years or less comprise 50% of the departing population, a career age that is likely to be during or shortly after the PhD training period. These proportions are largely consistent over time, and the large share of early career

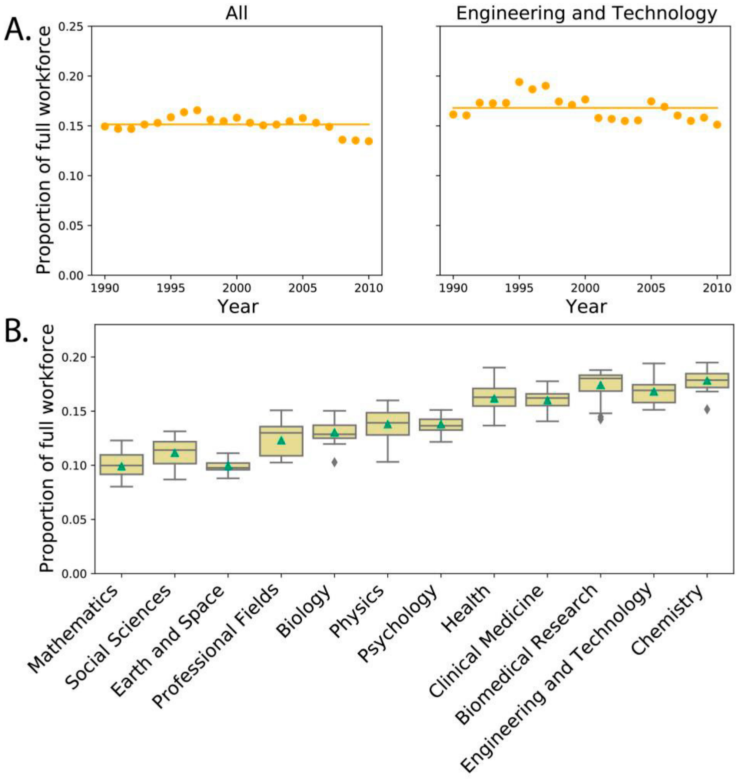


Figure 2A. The proportion of the total workforce who leave in each year, in all fields and in Science and Engineering, with the mean depicted as a flat line. B. The distributions of field specific turnover rates per year, with the mean as a triangle.

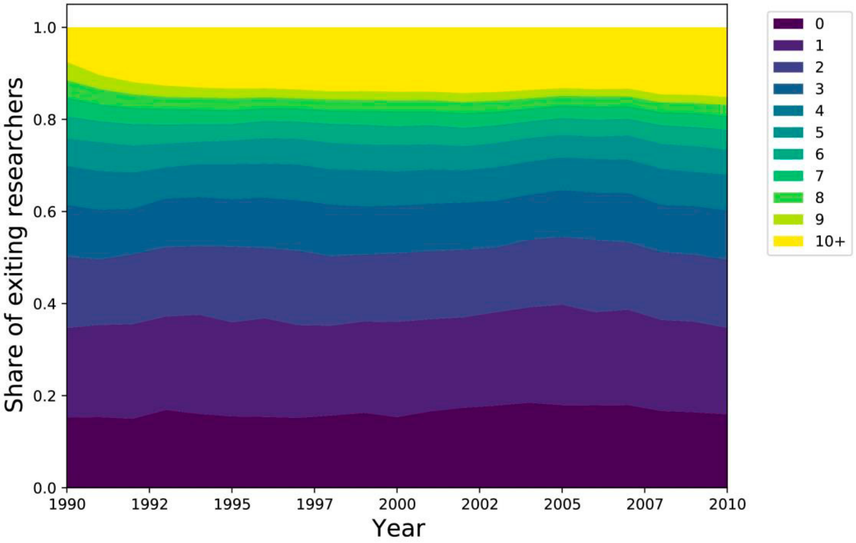


Figure 3. The share of researchers exiting academic science from each career age.

researchers suggests that the turnover within academic science occurs primarily in the scientific positions that require less research experience.

Discussion and Conclusion

Using an individual's range of publication dates, we have calculated the career ages of researchers in the academic workforce. Although there is an increasing number of researchers entering scientific fields each year, we find that there are commensurate increases in the number of researchers who leave academic science each year. We also find that as an average of 15% of researchers depart academic science each year, and these departing researchers tend to be in very early career stages. A frequent interpretation of the decreasing availability of long-term academic positions is that academic research has reached a crisis point where an oversupply of PhDs are trained for the number of available tenured positions (Cuthbert & Molla, 2015). These interpretations often rest on the traditional understanding of academic science in which tenured professors train students in order to fill similar tenure-track positions. However, Milojevic et al. have previously found that as research careers in academia shorten there may be a population of researchers whose roles are best characterized as temporary (Milojević et al., 2018). Our findings that most departures from academic science occur in the early career support the interpretation that the entry of scientists into PhD training may serve the shorter term goal of filling early career positions on research teams rather than the longer term goal of reproducing tenured faculty.

Furthermore, we find that the turnover rates and the share of departures at each career age have remained fairly stable over time between 1990 and 2010. Although academic researchers whose careers end involuntarily experience this as a scarcity of long-term academic positions, we argue that the prevalence of shorter careers within academic science may in fact be a larger condition of the working environment with team-based science. As such, the continuous entry of new researchers seems to be required to fulfil early career positions that are vacated either by researchers attaining sufficient experience for senior positions or departing academic science. In this way, it may be more accurate to argue that a larger proportion of early career roles are necessary to team science than ultimately can be accommodated in the long-term rather than a scarcity of tenure track positions.

We recognize that using the length of the publication career as a proxy for career age may be a limitation to our findings as there are field differences for when researchers typically publish their first paper and that researchers are publishing prior to earning their PhD more frequently now (Waaijer et al., 2016). In addition, we exclude researchers with only one publication, which may cause an underestimation of attrition rates. However, as both of these limitations should result in an underestimation of the extent to which scientific careers have been shortening, we feel confident that the interpretation of our results remains valid.

The high turnover rates within academic science are not reflective of an exodus of trained researchers from science because ample opportunities for scientific research are present outside academia as well. Although we are unable to identify the proportion of researchers who envision their academic work as training for industry careers, our finding that the fields with higher rates of turnover (Chemistry, Engineering and Technology, and Biomedical research) are also fields with strong ties to industry, suggests that scientists in these fields may be more likely to pursue industry positions after their training period. In this way, fields with a high probability of short careers among their participants could be interpreted as industry feeders. Researchers in industry are still active participants in scientific research, and with near 40% of doctorates entering large industry research firms by some estimates, industry hiring has been presented as a reliable way to transfer expertise from academia to the greater society (Zolas et al., 2015).

Voluntary departure for research opportunities in the government or private sector are common reasons for PhD recipients to leave academic research (NSF NCSES 2019). However, this does not invalidate the fact that many PhD trained researchers who desire to pursue long-term academic careers are stymied by a low availability of long-term tenure track positions (Bourne, 2013; Perlstein, 2014). Despite these factors, long term academic positions are still frequently framed by advisors as the preferred outcome of PhD training (Sauermann & Roach, 2012). These expectations and the nature of PhD training may need to be adjusted to reflect the reality that large proportions of scientists who participate in academic publishing during their training periods will not ultimately pursue long-term careers within academic science.

Acknowledgments

We thank the Sugimoto-Larivière joint meeting group and the Science Genome team for their helpful comments during meeting presentations. We also thank the ASIS&T SIG-MET participants for similarly helpful feedback. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-19-1-0391. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

References

Bourne, H. R. (2013). A fair deal for PhD students and postdocs. ELife, 2, e01139. https://doi.org/10.7554/eLife.01139

Cuthbert, D., & Molla, T. (2015). PhD crisis discourse: A critical approach to the framing of the problem and some Australian ‘solutions.’ Higher Education, 69(1), 33–53. https://doi.org/10.1007/s10734-014-9760-y

Holden, K. (2015). Lamenting the Golden Age: Love, Labour and Loss in the Collective Memory of Scientists. Science as Culture, 24(1), 24–45. https://doi.org/10.1080/09505431.2014.928678

Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. Scientometrics, 90(2), 463–481. https://doi.org/10.1007/s11192-011-0495-6

Larson, R. C., Ghaffarzadegan, N., & Xue, Y. (2014). Too Many PhD Graduates or Too Few Academic Job Openings: The Basic Reproductive Number R0 in Academia. Systems Research and Behavioral Science, 31(6), 745–750. https://doi.org/10.1002/sres.2210

Malmgren, R. D., Ottino, J. M., & Nunes Amaral, L. A. (2010). The role of mentorship in protégé performance. Nature, 465(7298), 622–626. https://doi.org/10.1038/nature09040

Milojević, S., Radicchi, F., & Walsh, J. P. (2018). Changing demographics of scientific careers: The rise of the temporary workforce. Proceedings of the National Academy of Sciences, 115(50), 12616. https://doi.org/10.1073/pnas.1800478115

National Science Foundation, National Center for Science and Engineering Statistics. 2019. Doctorate Recipients from U.S. Universities: 2018. Special Report NSF 20-301. Alexandria, VA. Available at https://ncses.nsf.gov/pubs/nsf20301/.

Perlstein, E. (2014, July 18). Generation Postdocalypse. Trade Secrets: A Blog from Nature Biotechnology. http://blogs.nature.com/tradesecrets/2014/07/18/generation-postdocalypse

Sauermann, H., & Roach, M. (2012). Science PhD Career Preferences: Levels, Changes, and Advisor Encouragement. PLOS ONE, 7(5), e36307. https://doi.org/10.1371/journal.pone.0036307

Waaijer, C. J. F., Macaluso, B., Sugimoto, C. R., & Larivière, V. (2016). Stability and Longevity in the Publication Careers of U.S. Doctorate Recipients. PLOS ONE, 11(4), e0154741. https://doi.org/10.1371/journal.pone.0154741

Walsh, J. P., & Lee, Y.-N. (2015). The bureaucratization of science. Research Policy, 44(8), 1584–1600. https://doi.org/10.1016/j.respol.2015.04.010

Zolas, N., Goldschlag, N., Jarmin, R., Stephan, P., Owen-Smith, J., Rosen, R. F., Allen, B. M., Weinberg, B. A., & Lane, J. I. (2015). Wrapping it up in a person: Examining employment and earnings outcomes for Ph.D. recipients. Science, 350(6266), 1367–1371. https://doi.org/10.1126/science.aac5949

Systematic characterisation and operationalisation of *cyber-physical convergence* in the context of Industry 4.0

Tausif Bordoloi[1,3], Philip Shapira[1, 2] and Paul Mativenga[3]

[1] *tausif.bordoloi@manchester.ac.uk; pshapira@manchester.ac.uk*
Manchester Institute of Innovation Research, Alliance Manchester Business School, University of Manchester,
Manchester, M15 6PB, UK

[2] *philip.shapira@pubpolicy.gatech.edu*
School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332-0345, USA

[3] *p.mativenga@manchester.ac.uk*
Department of Mechanical, Aerospace and Civil Engineering, University of Manchester,
Manchester, M13 9PL, UK

Abstract

Industry 4.0 is espoused by business leaders and policymakers as an innovation concept to capture value from *cyber-physical convergence*: the exploitation of manufacturing and industrial data to boost performance. In recent years, academics have labelled and promoted their works in terms of this concept, resulting in growth of the research domain. Yet precise definition and operationalisation of the notion of *cyber-physical convergence* itself has received little attention. We here describe a modular text-mining approach that builds on the technological underpinnings of Industry 4.0 to operationalise *convergence*. In doing so, we identify a list of relevant keywords that corrspond to five data-centric capabilities: Monitoring, Analytics, Modelling, Transmission and Security. These capabilities set the stage for further bibliometric work to systematically delineate the *convergence* research domain and track its growth, as it interacts with Industry 4.0.

Introduction

In the early 2010, the German Government coined the innovation policy concept “Industrie 4.0” (also known as Industry 4.0”) to capture value from Germany’s manufacturing and industrial sectors by championing *cyber-physical convergence* – the combination of the cyber (as embodied by the Internet) and physical worlds (Kagermann, Wahlster, & Helbig, 2013). A key focus of Industry 4.0 is to influence the production of scientific knowledge (Plattform-Industrie-4.0, 2019, p. 41). Indeed, policy interest has been accompanied by growth in academic research worldwide, where Industry 4.0 has been used to label and promote, among other things, various types of advanced technologies (e.g., augmented reality (Masood & Egger, 2019)), production models (e.g., lean production (Mrugalska & Wyrwicka, 2017)) and business models (e.g. sustainability (de Man & Strandhagen, 2017)). These apparent policy impacts, however, must be viewed against a literature where little or no effort has been made to systematically define *convergence* and delineate the research domain that Industry 4.0 seeks to influence. It is no surprise, therefore, that attempts by the bibliometric community to evaluate scientific output associated with Industry 4.0 have tended to be *ad hoc* and imprecise in scope (see, for example, Muhuri, Shukla, and Abraham (2019)). This reflects both a conceptual and boundary gap, ignoring which will limit effective measurement of research performance indicators, and hamper funding allocations (Boswell & Smith, 2017).

In this paper, we contribute definitionally and methodologically. We examine the technological underpinnings of Industry 4.0 as elucidated by its original proponents, and characterise *cyber-physical convergence* in a manner that closely aligns with these underpinnings. Our characterisation focuses at the level of manufacturing and industrial data. With this approach we separate random signifiers associated with the concept. Next, we operationalise the definition using natural language processing methods. In doing so, we identify five data-centric capabilities: (i) Monitoring, (ii) Analytics, (iii) Modelling, (iv) Transmission and (v) Security. These capabilities collectively specify, for the first time to our knowledge, a boundary that can delineate the *cyber-physical convergence* research domain.

In the next section, we discuss the technological considerations of Industry 4.0, based on which we characterise *convergence*. In Section 3, the definition is operationalised and capabilities are identified. Section 4 presents our discussions and concluding remarks.

Technological underpinnings of Industry 4.0

We identified and reviewed a set of foundational policy and technical literature (Geisberger & Broy, 2015; Hellinger & Seeger, 2011; Kagermann, Wahlster, & Helbig, 2013) (foundational because this literature frames Industry 4.0 through the lens of their original proponents) to enable understanding of the technological foundations of *cyber-physical convergence*. Table 1 presents a definition of *convergence* according to the German National Academy of Science and Engineering (or Acatech) that has played a central role in establishing and driving the Industry 4.0 economic, social and technological narrative in Germany.

As evident from the definition, data is the fundamental driver of *cyber-physical convergence*. The definition underscores a set of data workflows, whereby data is first sensed (or collected) from machines. The collected data is then subjected to a series of processes, which culminates in the generation of actionable insights. Clearly, this definition does not equate *cyber-physical convergence* with new advanced technologies. Advanced technologies such as augmented reality will not necessarily produce a *convergence* phenomenon unless the data that they produce is collected, processed and then applied to drive strategic decisions. Conversely, even legacy machinery can be classified as being convergent if their data is effectively captured and monetised (Kagermann, Wahlster, & Helbig, 2013). This definition is useful in that it provides a way to separate out random signifiers for the purpose of operationalising it.

Table 1. Definition of *cyber-physical convergence* according to Industry 4.0

Concept	*Definition (according to Acatech)*
Industrie 4.0 Geisberger and Broy, (2015, p. 64)	*Convergence* involves the ability of manufacturing systems to: (a) capture physical data from the environment in parallel via sensors and to merge and process this data – and to do so both locally and globally and in real time; (b) use the information that they have gathered to interpret the situation in terms of predefined goals; (c) detect, interpret, deduce and forecast malfunctions, problems and threats; (d) integrate, regulate, control and interact with components and functions; and (e) carry out globally distributed and networked control and regulation in real time.

Operationalising *cyber-physical convergence*

To operationalise the above definition of *cyber-physical convergence*, there is a more basic question to address first – how exactly do researchers working in the domain use the term "data" in their publications? This question is pertinent because it concerns translating the data-centric capabilities of Table 1 into a controlled lexicon of keywords.

Building a lexicon of relevant keywords

Figure 1 shows our overall lexicon building methodology. We extract benchmark publications from the scientific publication database Scopus and analyse them with natural language processing tools to produce an initial set of mid- and high-frequency keywords. Capturing high-frequency keywords from publications is an effective means to operationalise emerging technological domains (Shapira, Kwon, & Youtie, 2017). Mid-frequency keywords play a complementary role in this process because they reveal not-too-common words in a specific field (Luhn, 1958). The relevance of these keywords with respect to the publication corpus is determined by applying a statistical measure called term frequency-inverse document frequency (TF-IDF) (Chen & Xiao, 2016), and also by checking their meaningfulness in regards to specific word combinations (phrases). Our analysis finds a set of *convergence*-specific keywords. Next, following a modular methodology described by Mogoutov and Kahane (2007), we test these keywords for their specificity to obtain additional, complementary keywords, which are then added to the list of relevant keywords.

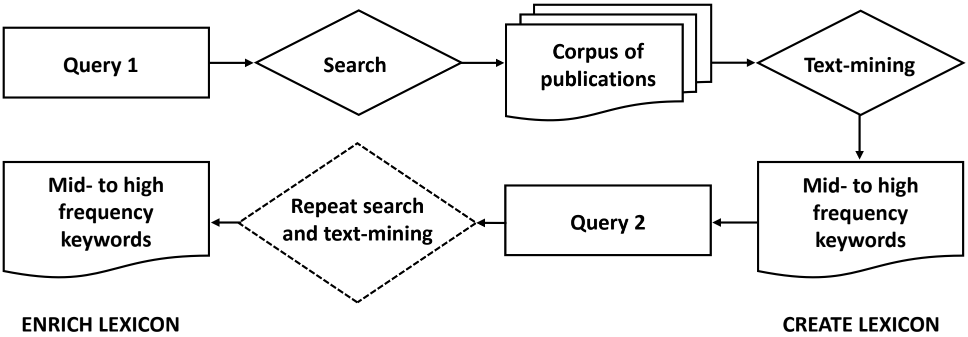

Query 1
Search
Corpus of publications
Text-mining
Mid- to high frequency keywords
Repeat search and text-mining
Query 2
Mid- to high frequency keywords
ENRICH LEXICON
CREATE LEXICON

Figure 1: Modular lexicon building methodology. See text for details.

S1. Retrieving benchmark publications

We used a first strategy **Query 1** (Title = data AND ("industrie 4.0" OR "smart manufacturing" OR "industrial internet")) in Scopus. We selected journal and conference articles in English and German languages from 2010 to 2020 (Scopus provides English translation of a limited number of German publications). This corpus was built in June 2020. Our search returned 118 publications, consisting of all available publications that demarcated these concepts using "data". We then manually checked the titles, abstracts and keywords of the publications, and removed irrelevant ones. The final cleaned dataset contains 96 publications.

S2. Identifying relevant keywords

The titles, abstracts and keywords of the cleaned dataset were analysed using the content analysis tool QDA Miner in combination with its sister text-mining module WORDSTAT. QDA Miner provides various statistics such as TF-IDF, term frequency and the number of publications in which terms are found. We extracted individual words and phrases along with their statistical measures. The raw keyword set contained 300 unique keywords. These were manually screened to remove nonsensical/trivial terms, resulting in a parent set of 77 keywords. These keywords were then subjected to additional processing as follows. First, complementary terms related to the concepts in the parent set – "industry 4.0", "cyber-physical", "iiot", "industrial iot" – were grouped together. To these, we added the terms "factory" and "shop floor" to capture papers that express the same interpretation of *convergence*, even though these papers may not explicitly use the concepts in their title, abstract or keywords. Combining all these terms expands our search string to ("industrie 4.0" OR "industry 4.0" OR "smart manufacturing" OR "cyber manufacturing" OR industrial internet" OR "iiot" OR "industrial iot" OR "cyber physical" OR "cyberphysical" OR "factory" OR "shop floor"). This string **(Query 2)** represents a reasonably robust search space from which papers that are judged to embody *convergence* can be extracted. Second, we selected terms to operationalise *convergence*. We classified the remaining keywords in the parent set according to their frequencies (ranging from 6 to 65) and TF-IDF scores (between 8 and 48). These values were calculated in the aforementioned text mining process for all terms in all of the documents in the corpus. We selected terms with mid-to-high frequencies (equal to or greater than 30) and TF-IDF weights (equal to or greater than 24) because they were judged to be associated with key research topics, and they produced highly pronounced relevance signals, indicating their importance within their respective publications. Based on these criteria, eight keywords were selected: "monitoring", "sensing", "analytics", "modelling", "transmission", "network", "communication" and "security" (Table 2).

Table 2. Capabilities with relevant keywords and phrases

Capability	*Relevant keyword*	*Relevant phrases (Examples)*	*No. of publications with this keyword*	*Total frequency of the keyword*	*TF-IDF score of the keyword*
Monitoring	Monitoring	data monitoring	12	31	27.9
	Sensing	data sensing, sensor data	14	33	27.4
Analytics	Analytics	data analytics, industrial analytics	24	65	47.8
Modelling	Modelling	data modelling	20	42	28.4
Transmission	Transmission	data transmission, real-time transmission	10	35	34.2
	Network	software-defined networking	22	51	32.4
	Communication	data communication	16	32	24.8
Security	Security	cybersecurity, security in Industrie 4.0	15	38	30.5

Each of these eight keywords represents a specific data-centric capability. The meanings of these keywords were then examined in relation to specific phrases, and also by manually checking titles and abstracts. Association between specific words mean that a distinct aspect of the domain is being discussed. Thus, phrases like "data monitoring", "sensing data", "data acquisition", "data analytics", "data modelling", "data transmission" and "data communication" (Table 2) helped place the keywords in the specified context, and categorise them into five capabilities: Monitoring, Analytics, Modelling, Transmission and Security.

S3. Enriching the lexicon at capability levels

To enrich the lexicon with relevant keywords, we pursued a modular query enrichment methodology similar to Mogoutov and Kahane (2007), whereby subfield keywords were extracted in another round of querying in Scopus (using Query 2) and then tested using QDA Miner (as discussed above) for inclusion into the lexicon. Table 3 shows a total of 16 keywords of the enriched lexicon that can collectively delineate the *cyber-physical convergence* research domain (additional relevant keywords extracted in this step are marked in bold) into five data-centric capabilities (Note: Although it has a negative connotation, "attack" was identified as relevant and popular in our text-mining approach because researchers use it in different security-related contexts: "to block an attack", "encryption-based cyber-attack" or "securing machines against attacks").

Table 3. Enriched lexicon to operationalise and delineate *cyber-physical convergence*

Capabilities	*Brief description of capabilities*	*Scopus search query for step S3*	*Keywords included in the final lexicon*
Monitoring	Collection of data from manufacuring and industrial processes	(TITLE ("monitor*" OR "sens*") AND TITLE ABS (Query 2))	monitoring, sensing, **acquisition, collection**
Analytics	Analysis of the collected data to extract useful information	(TITLE ("analytics") AND TITLE ABS (Query 2))	analytics, **learning, neural network, intelligence**
Modelling	Virtual modelling of the states, conditions and behaviours of systems	(TITLE ("model*") AND TITLE ABS (Query 2))	modelling, **simulation**
Transmission	Transmission of data across different software and hardware systems	(TITLE ("transmi*" OR "network*" OR "communicat*") AND TITLE ABS (Query 2))	transmission, network, communication, **wireless**
Security	Cybersecurity to safeguard critical manufacturing assets	(TITLE ("secur*") AND TITLE ABS (Query 2))	security, **attack**

Discussion and conclusions

This paper presents a text-mining approach that builds on the technological underpinnings of Industry 4.0 to define *cyber-physical convergence*. This definition is then operationalised using a modular query building process that allows us to characterise *convergence* into five data-centric capabilities (corresponding to a total of 16 relevant keywords).

We recognise that our text-mining approach has limitations. There may be differences regarding what other experts and stakeholders consider important as to the definition of *cyber-physical convergence*. This may affect the choice of keywords that we have identified. Nonetheless, the fact remains that our attempt to closely align with the Industry 4.0 definition of *cyber-physical convergence* will allow us to separate out signifiers and delineate the research domain more precisely.

Our attempt in this paper forms part of a broader research to investigate policy interactions with academic research growth and evolution. In the next stages of this work, we will apply these keywords (with appropriate inclusion and exclusion terms) to perform a bibliometric search of the *cyber-physical convergence* research domain.

References

Boswell, C., & Smith, K. (2017). Rethinking policy 'impact': four models of research-policy relations. *Palgrave Communications, 3*(1), 1-10.

Chen, G., & Xiao, L. (2016). Selecting Publication Keywords for Domain Analysis in Bibliometrics: A Comparison of Three Methods. *Journal of Informetrics, 10*(1), 212-223.

de Man, J. C., & Strandhagen, J. O. (2017). An Industry 4.0 Research Agenda for Sustainable Business Models. *Procedia CIRP Conference on Manufacturing Systems, 63*, 721-726.

Geisberger, E., & Broy, M. (2015). *Living in a Networked World: Integrated Research Agenda Cyber-Physical Systems (agendaCPS)*. Acatech.

Hellinger, A., & Seeger, H. (2011). *Cyber-Physical Systems. Driving Force for Innovation in Mobility, Health, Energy and Production*. Acatech.

Kagermann, H., Wahlster, W., & Helbig, J. (2013). *Recommendations for Implementing the Strategic Initiative INDUSTRIE 4.0: Securing the Future of German Manufacturing Industry*. Acatech and Forschungsunion.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development, 2*(2), 159-165.

Masood, T., & Egger, J. (2019). Augmented reality in support of Industry 4.0 - Implementation challenges and success factors. *Robotics and Computer-Integrated Manufacturing*, 58, 181-195.

Mogoutov, A., & Kahane, B. (2007). Data Search Strategy for Science and Technology Emergence: A Scalable and Evolutionary Query for Nanotechnology Tracking. *Research Policy*, 36, 893-903.

Mrugalska, B., & Wyrwicka, M. K. (2017). Towards Lean Production in Industry 4.0. *Procedia Engineering, 182*, 466-473.

Muhuri, P. K., Shukla, A. K., & Abraham, A. (2019). Industry 4.0: A Bibliometric Analysis and Detailed Overview. *Engineering Applications of Artificial Intelligence*, 78, 218-235.

Plattform-Industrie-4.0. (2019). *2019 Progress Report*. Federal Ministry for Economic Affairs and Energy (BMWi).

Shapira, P., Kwon, S., & Youtie, J. (2017). Tracking the Emergence of Synthetic Biology. *Scientometrics*, 112(3), 1439-1469.

Beyond the Web of Science: an overview of Brazilian papers indexed by regionally relevant databases

André Brasil[1]

[1] *a.l.brasil@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies (CWTS)
Kolffpad 1, 2333 BN Leiden (The Netherlands)

Abstract

The Brazilian evaluation system of research and graduate education is under revision. For scientific publishing, policymakers and evaluators consider a more standardised assessment, trading some current qualitative aspects for expanded use of journal-based indicators from international databases such as the Web of Science. This paper aims to provide evidence for a better-informed discussion, analysing the complete data set of Brazilian articles published from 2013-2018 regarding: i) coverage by the WoS and regionally relevant databases (e.g., Latindex, SciELO); ii) incidence of the local language in the country's publications, and the impact it has on coverage by regional and international databases; iii) disciplinary variations and disparities in thematic coverage across languages and databases. Results show half of the Brazilian article output is not found in the WoS, and the normalised distribution of indexed publications across disciplines is hugely unbalanced. Publications not in WoS are predominantly in Portuguese, with a significant share indexed by regional databases, often addressing topics not covered in WoS. The main conclusion is that Brazilian science goes beyond WoS, and evaluators should strive for a sound and comprehensive assessment to capture its complexity, instead of trading it for restraining, short-sighted simplicity.

Introduction

Brazil counts with a national evaluation system which is conducted by the Brazilian Agency for Support and Evaluation of Graduate Education (CAPES). As the name of the agency in charge suggests, such evaluation focuses on masters and doctoral courses, but the Brazilian Society for the Advancement of Science (SBPC) indicates over 95% of all Science & Technology research conducted in the country comes from these graduate courses. Thus, it is possible to state that evaluating graduate education is nearly the same as assessing the whole science system in Brazil (Nobre et al., 2017; SBPC & ABC, 2020).

First conceived in the 1970s, the evaluation of graduate programs was initially intended to better allocate funding for the development of science in the country. Over time, the model grew in size and complexity, becoming a high-stakes assessment of almost 8000 master's and doctoral courses organised into nearly 5000 research units or programs (PPG). Every four years, all PPG in the country go through a compulsory, government-funded evaluation that grades them on a scale from '1' through '7'; the last representing the highest level of excellence. Positive results not only guarantee status for the PPG and the institutions that promote them, but also implicate in an increase in funding, a higher number of available scholarships, and access to a wider range of grants. In comparison, a low performance will not only lead to funding cuts, but it can also threaten the programs' existence. For instance, research units graded '1' or '2' are no longer allowed to enrol new graduate students and must suspend their activities completely as soon as the last of the currently enrolled students graduate (CAPES, 2017; Nobre et al., 2017).

While an additional study is being conducted to explore the complexity of the national evaluation system in Brazil, for this paper it is relevant to know that the actual evaluation relies heavily on mixed methods. Disciplinary and interdisciplinary panels from 49 distinct evaluation fields perform the assessment of PPG with qualitative and quantitative data collected yearly from every research unit in the country. From the available evidence, several dimensions are assessed by the committees in charge, from infrastructure to educational results and societal impact of the PPG. Scientific production is, as expected, one of the most valued dimensions in the process, including specific evaluation of distinct types of output such as books, technical and artistic products, conference proceedings, and papers, among others.

The assessment of papers published by graduate programs in the country follows the principles of the broader evaluation system: each of the 49 research areas receives a comprehensive list of each PPG's publications, together with a series of indicators gathered and calculated from national and international databases. Panels composed by established researchers in each area interpret the available information – combined with their inherent knowledge in the field – and generate what is known as Qualis Journals: a nine-level classification system for all journals used by Brazilian researchers during the evaluation cycle (Barata, 2016; CAPES, 2021b).

Qualis has been an essential element in Brazilian evaluation since 1998. The classification system evolves over time to reflect the advancements in scientific publishing as well as in information systems (Barata, 2016). To promote changes in the current evaluation cycle (2017–2020), CAPES established a working group to review Qualis procedures (CAPES, 2018). In a preliminary report, the group states that the new assessment model must induce internationalisation of both publishing and journal indexing. The initiative proposes to reduce the qualitative perspective, attributing to journal-level indicators – such as *jif* and *h5* – most of the responsibility for the final classification (dos Santos et al., 2019).

Considering the evaluation literature strongly discourages practices which rely heavily on indicators to assess scientific productivity, especially those dependent on journal-level metrics (Hicks et al., 2015), it is not surprising that a consensus has not yet been found among the research-area representatives working with CAPES. Several disciplines, especially those in the Social Sciences and Humanities (SSH), are concerned with the proposal for a 'New Qualis', especially because the current focus is on adding weight to the value of publications indexed by databases such as the Web of Science (WoS) and Scopus, while undervaluing the ones indexed by regional databases, mostly in local language. As a consequence, even though ordinance no. 150 (CAPES, 2018) established a three-month deadline for publication of a detailed report by the Qualis Journals working group, almost three years have passed, and the document remains unpublished.

The main goal of this research is to contribute to this current debate around the 'New Qualis', primarily by investigating topics that may shed light over concerns such as: i) How is the coverage of Brazilian papers in international databases? ii) How representative is the local language for the country's publications, and how different is the coverage from regional to international databases? iii) In case significant differences are found, how extreme are the disciplinary variations? iv) From a thematic standpoint, are there research topics that have more space in regional databases?

Methods and data

The central data set that supports the current research comes from all Brazilian graduate programs. It is a virtually complete list of all publications produced in these programs, by both faculty members and student body alike. There are two main reasons why this data set covers the majority of the actual output from graduate programs. The first one lies on the aforementioned high stakes of evaluation. Since performance relates to funding and to the continued existence of research programs, their directors, university pro-rectors, and all PPG researchers are concerned with the quality of the data provided. The second reason is that, for many years, CAPES has counted on information systems to collect PPG data to perform the evaluation. Distinct systems have been developed over time and the data collection is now conducted through the Sucupira Platform: an integrated system that is robust enough to deal with the size of the Brazilian National System of Research and Graduate Education (SNPG). The platform is not only open continuously for data submission, but it grants the general public direct and real-time access to all of it. This means all PPG researchers and stakeholders become part of a relevant auditing system. As such data is subsequently audited by the committees in charge of the evaluation, there is also an authenticity layer of control (CAPES, 2021b; Siqueira, 2019).

Integrated data sets from graduate education in Brazil are made available through CAPES' Open Data platform (CAPES, 2021a). In this research, the R language (R Core Team, 2020) was used to combine and clean eight of the available data sets which relate to four distinct categories in the database: i) general information from graduate programs ii) scientific output from graduate programs iii) detailing of bibliographic production from PPG iv) authorship details of papers and reviews. The most recent Qualis rankings were subsequently gathered from the Sucupira Platform (CAPES, 2021b) and combined with the broader data set.

The resulting data include all papers published from 2013 to 2018, totalling more than 1,3 million records of around 750 thousand individual publications. The reason for the difference is that CAPES' records are PPG based, which means publications co-authored by researchers from distinct research programs will be recorded for each individual PPG. Publications from 2019/2020 were not yet included in the study because data for those years are only partially available, with relevant details such as DOI and journal ISSN still missing (CAPES, 2021a).

Once the work on the core data was completed, a series of complementary databases were consulted to enrich the resulting data set. For that, key fields like DOI, ISSN, and additional information on author last names, journal volumes/issues, page numbers and titles of publications were all used to match the consolidated data with the following additional sources:

i) SciELO – bibliographic database structured on a cooperative model of Open Access (OA) journals, which are selected based on a set of quality and operational criteria. The network focuses on the scientific communication needs of developing countries, especially those in Latin America and the Caribbean. It was first established in Brazil and now extends to 15 other countries, including Portugal, Spain, and South Africa (SciELO, 2021).

ii) RedALyC – similar to SciELO, it is a network of OA journals focused on scientific output from Latin America and the Caribbean, while also extending to other Portuguese and Spanish speaking countries. It reaches 26 countries and establishes a series of criteria to include journals in its database (RedALyC, 2021).

iii) Latindex – another network focused on OA publishing. It reaches 23 countries, mostly Spanish or Portuguese speaking ones, counting with a directory of 29.192 indexed journals. Even though the broader coverage may be quite useful to map the Brazilian scientific output, only a small percentage of these journals have been through the compliance process with the Latindex methodology, which includes criteria similar to those adopted by the other two networks mentioned. These quality-assured journals form a Latindex subset of 2265 journals, known as Catalogue 2.0 (Latindex, 2021).

The enriched data set created in this research allowed a series of analyses on coverage of the selected regional databases, publication languages, research areas, and more. Nevertheless, at least one international database was also needed, and the Web of Science (WoS) was chosen to complement the research. According to Chavarro et al. (2018), this database is one of the main data sources used to obtain bibliographic data for many quantitative research assessments, and this includes the impact factors which inform the Brazilian Qualis classification of journals.

Findings and discussion

From 2013 to 2018, CAPES' database of scientific production lists a total output of 752.453 papers and reviews, published in 29.679 distinct journals (merging ISSN for electronic and print versions). All of them were classified under the Qualis criteria which, in the current scale, consists of grades A1, A2, A3, A4, B1, B2, B3, B4, and C/NP (A1 being the top grade). Each disciplinary committee determines the score attributed to each of the possible grades, and individual publications receive corresponding points during the evaluation process, except C/NP. This rank is reserved for low quality or predatory journals, as well as those which the committees consider lacking scientific rigour (e.g., non-peer-reviewed publications) (CAPES, 2021b).

To best represent the universe of valid journals, all results in this paper exclude those which were ranked C/NP, as well as any others which may have been disqualified by the evaluators during the data auditing process. As a consequence, the number of considered papers drops to 585.945, published in 23.508 journals. The filtered CAPES database was then matched with the Web of Science (WoS), following the guidelines described by Visser et al. (2021). A total of 9.597 journals (40,8%) were found in the database, with 298.170 papers published (49,4%). Figure 1 shows the results of this analysis.

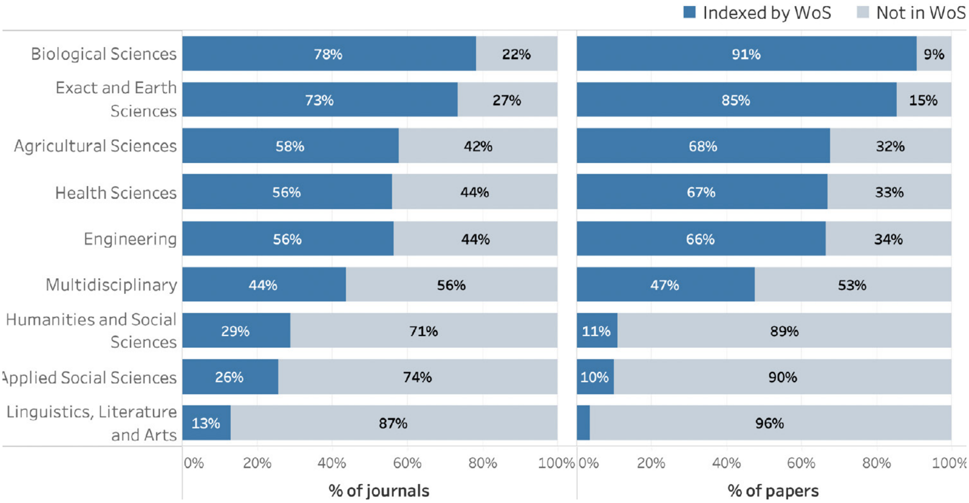


Figure 1. Percentage of journals and papers indexed by WoS (2013 - 2018), according to research areas adopted by CAPES

As Figure 1 shows, there is a considerable difference in WoS coverage when results are grouped according to the nine research areas defined by CAPES: a meso-level aggregation of the 49 evaluation fields conducting the evaluation process. As it can be seen, 78% of the journals used by Biological Sciences PPG to publish their papers (2013–2018) are indexed by the WoS. In contrast, the same is true for only 13% of those from Linguistics, Literature, and Arts.

From the papers' perspective, it can be noted that the number of publications in research areas with broader WoS coverage tends to be even higher. For instance, 73% of Exact and Earth Sciences journals are indexed by WoS, and this number rises to 85% when looking at the paper level. Opposite to that are all Social Sciences and Humanities (SSH), where the percentage of papers in the Web of Science is even smaller than their journal coverage, indicating a lower-than-average number of publications in such journals.

Of course, caution should be used to interpret disciplinary differences observed in Figure 1. Scholars such as Tijssen et al. (2006) and Chavarro et al. (2018) have already highlighted the dangers in analysing differences in database coverage across research areas, as large variations are often influenced by internal biases in the database's own coverage. Nevertheless, for the trained eye, Figure 1 seems to paint a very extreme picture.

To help analyse if the disciplinary variation in the Web of Science coverage is more prominent for Brazilian scientific publications, Figure 2 was designed from data collected from the bibliometric version of the WoS core collection hosted at the Centre for Science and Technology Studies (CWTS). All papers and reviews from the 2013-2018 period were included in the visualisation designed using the VOSviewer software (van Eck et al., 2009), and the publications were grouped according to the micro-level field classification (Waltman et al., 2012).

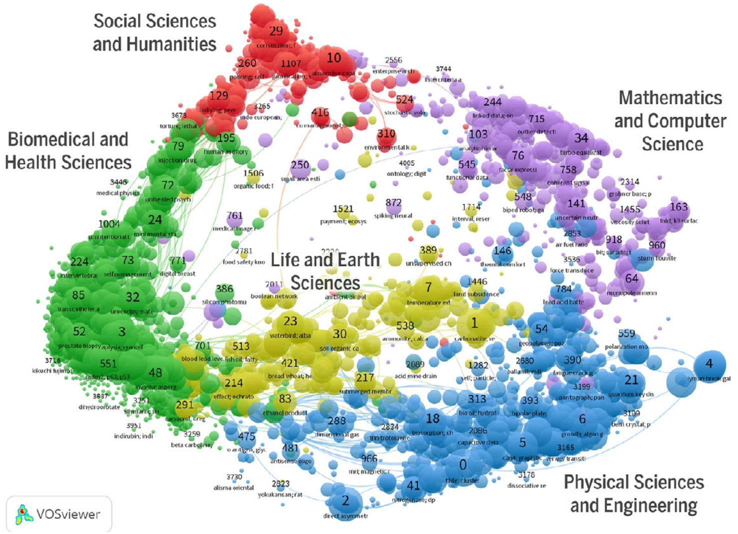


Figure 2: Scientific production indexed by the Web of Science (2013-2018), mapped according to CWTS micro-level field classification

Figure 2 shows 4013 clusters of publications in the WoS database. The size of the circles represents the number of publications in each cluster and their positions are calculated from the citation traffic between all papers. Five main fields of science are shown, and they confirm that there is a proportionally smaller number of SSH clusters and publications in the database. To see how Brazilian scientific publishing relates to the broader WoS coverage, Figure 3 maintains the configuration from the previous map but including only Brazilian papers and reviews, with resized circles according to the proportion of the country's available publications.

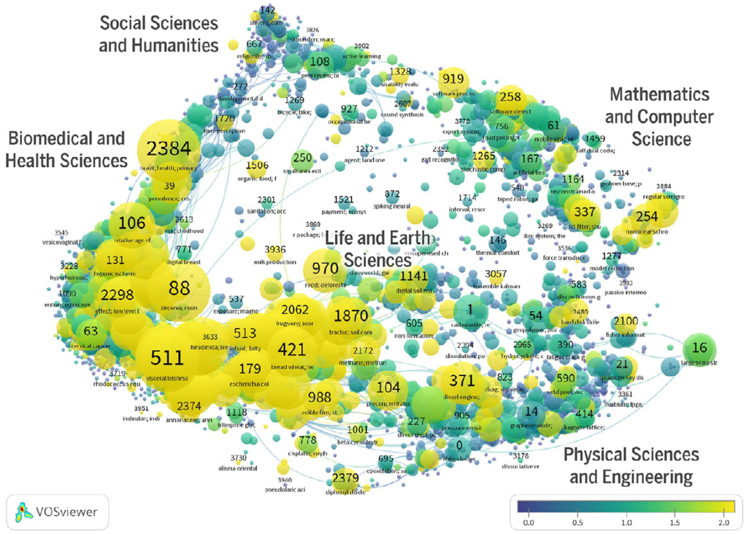


Figure 3: Brazilian scientific production indexed by the WoS (2013-2018), mapped according to CWTS micro-level field classification

The size variation of clusters in Figure 3 is evidently more pronounced than in the previous map, and the colour overlay helps understand the significance of that. Cluster colours pointing to 1.0 in the included scale imply the proportion of Brazilian publications is around the same as for the global database. Lighter colours mean Brazil produces a higher proportion in these fields, and the darker colours mean the exact opposite. The resulting visualisation shows Brazilian publishing is proportionally higher across the Biomedical and Health Sciences, and in some Life and Earth Sciences clusters. Isolated points of higher relative productivity can be found in other research areas, but most other clusters are underrepresented.

The main message from Figure 3 seems to be that Brazilian publishing behaviour within the Web of Science amplifies the coverage biases already observed in the whole database. To confirm this perception, the internal coverage calculation described by Moed and Visser (2007) will be applied. This indicator analyses the extent to which the articles included in the WoS cite other articles which may or may not be found in that same database. From the number of missing references, it is possible to infer the percentage of scientific publications that are not indexed by the Web of Science. Figure 4 shows the results of the analysis but grouping the CWTS field classification systems in their macro-levels, for clarity purposes.

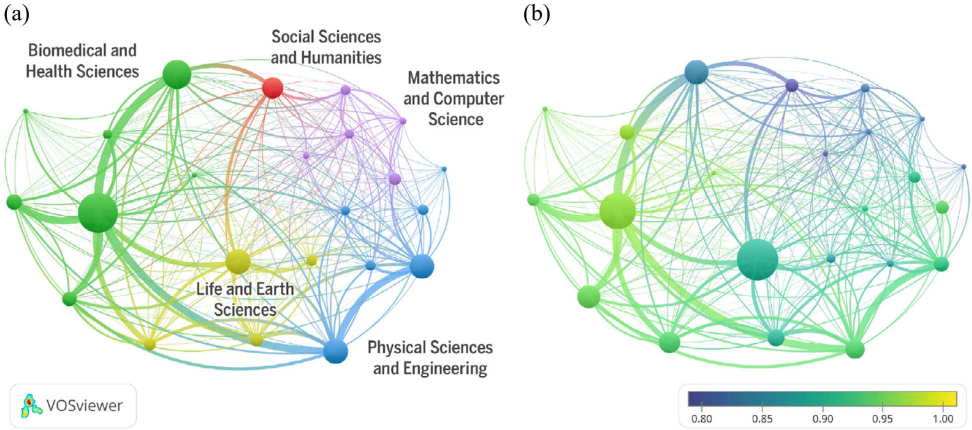

(a)
(b)
Biomedical and Health Sciences
Social Sciences and Humanities
Mathematics and Computer Science
Life and Earth Sciences
Physical Sciences and Engineering
VOSviewer
0.80
0.85
0.90
0.95
1.00

Figure 4: Scientific production mapped according to CWTS macro-level field classification for (a) complete WoS database (2013-2018), and (b) only Brazilian publications, with relative internal coverage overlay

Figure 4a shows the 23 broad disciplines of CWTS macro-level field classification. The colour clustering follows the same scheme seen in Figure 2, and the size of each circle indicates the number of papers in the WoS database (2013–2018). To design Figure 4b, the data was filtered to show only the Brazilian publications with proportionately resized clusters. As in the previous VOSviewer maps included in this paper, it is possible to see differences in the publication distribution across fields from (a) to (b). For instance, Brazil tends to produce relatively more in most Life and Earth Sciences clusters, but much less in the single SSH cluster.

To design the colour overlay seen on Figure 4b, the average internal coverage (Moed & Visser, 2007) was calculated for each of the 23 broad disciplines in both data sets: the complete one seen in (a) and the Brazilian one displayed in (b). The results show, for example, that internal coverage in (a) for the largest Biomedical and Health Sciences cluster is of 91%, for the most prominent Life and Earth Sciences is of 73%, while for the single SSH cluster is of only 34%. The numbers for Brazil are slightly lower in (b), at 88%, 65%, and 27%, respectively. This means Brazilian publications are citing a larger percentage of works not indexed by the WoS.

A second step for the colour overlay on Figure 4b was to calculate the ratio between the broader database internal coverage and the Brazilian one. For SSH, for instance, this would mean 0,27 / 0,34, indicating Brazil's relative coverage is of 79% of the outcome seen for the unfiltered database. The results displayed in the overlay show Brazilian internal coverage is lower across all 23 broad disciplines, and the largest variations are seen exactly on clusters that already registered low coverage in the broader database. This analysis confirms the original impression that coverage biases in the Web of Science are exacerbated in Brazilian output. Is this a result of publishing barriers or a matter of choice to publish in journals indexed elsewhere?

Maybe the answer can come partially from scholars such as Gibbs (1995) and Tijssen et al. (2006), who discussed challenges that researchers in the developing world need to face to overcome publishing barriers. Among them are both disciplinary and language obstacles, as English journals dominate databases such as the Web of Science, with much less room for Portuguese or Spanish publications. This is a significant problem for Brazil, as the most recent Education First Proficiency Index (2020) – which ranks 100 regions according to English Skills – places the country in the low proficiency group (53rd position). Besides that, the British Council (2014) reports only 5% of Brazilians have some knowledge of English, while only ⅕ of those present some level of fluency. The numbers are also not good, yet less abysmal, when researcher population is considered. According to Vasconcelos (2008), who analysed data from the *curricula vitae* of over 50 thousand Brazilian scientists, around 33% of them declared to be proficient in English. Even disregarding biases from the self-declaration of proficiency seen in the CVs, still only a third of Brazilian researchers would not find language an obstacle to publish in English. Considering these language limitations, Figure 5 shows all Brazilian papers from the 2013–2018 period, isolating the publication percentages according to WoS coverage and the main language reported for the individual papers.

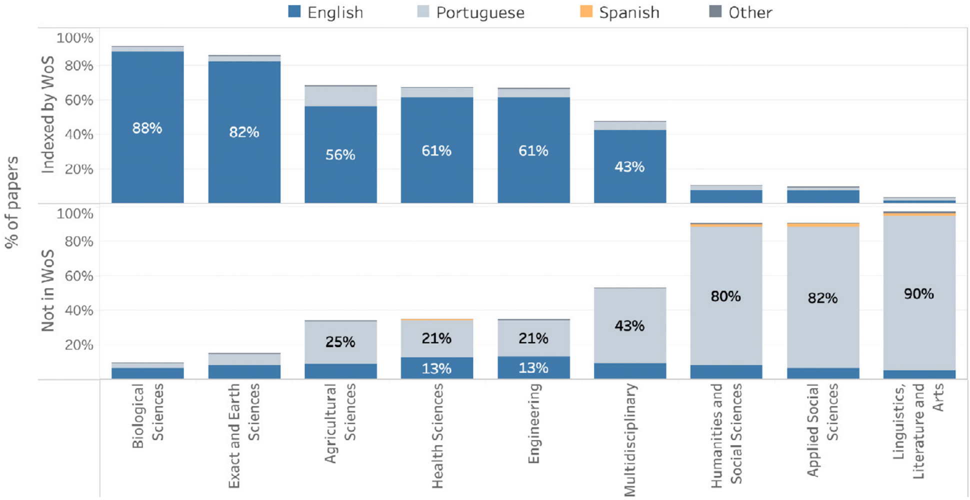


Figure 5: Comparison of language profile of Brazilian papers (2013-2018), in and out of WoS, according to research areas adopted by CAPES

Disciplinary differences are, once again, strikingly evident in Figure 5. Most of the Brazilian papers indexed by the Web of Science are in English, and publications not covered are mostly in Portuguese. SSH papers in local languages account for more than 80% of publications in all three research areas. The Multidisciplinary field is also impressive as publications are nearly split in the middle regarding WoS coverage, with language participation mirroring each other.

Figure 5 also shows that there is only a very small share of papers in Spanish or other languages, which add to less than eight thousand publications (a little more than 1% of the total).

After identifying that most papers not indexed by the Web of Science are in Portuguese, even in research fields such as Health Sciences and Engineering, we may wonder if such publications would be found in regionally relevant databases. Figure 6 shows the results of such analysis by mapping the whole set of Brazilian publications (2013–2018) with the Latindex, SciELO, and RedALyC databases.

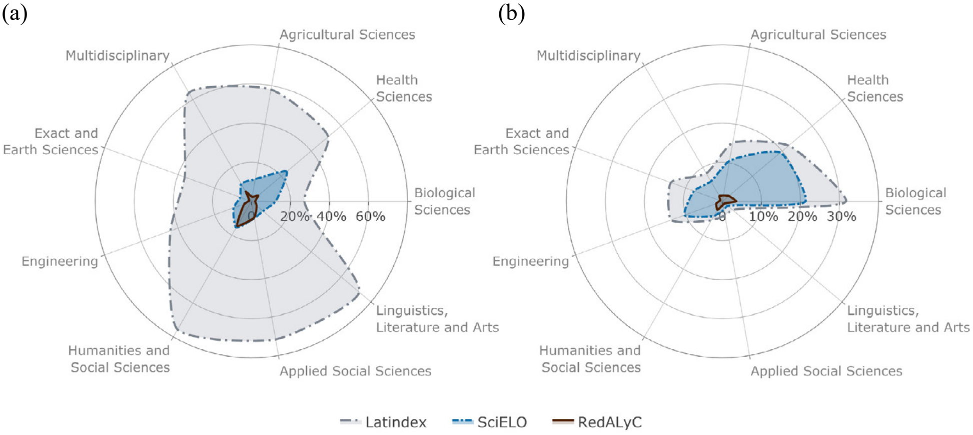


Figure 6: Brazilian papers (2013-2018) not indexed by the WoS but available at Latindex, SciELO, or RedALyC in (a) Non-English and (b) English

The two polar charts included in Figure 6 show the percentage of publications in each of the CAPES-adopted research areas that can be found in Latindex, SciELO, and RedALyC. The left chart shows non-English publications, which we have seen are mostly in Portuguese. The broader Latindex directory covers a significant percentage of papers not found in the WoS, reaching nearly 80% in the Humanities and Social Sciences, with its lowest coverage in Biological Sciences, where it is of around 30%. SciELO and RedALyC, counting with more curated collections, cover a smaller share, but over 20% of the 'missing' Health Sciences papers are included in SciELO, for example. For English publications not included in the WoS collection, Figure 6b shows that there is very little to find in RedALyC, but Biological and Health Sciences have a significant share of papers in SciELO and Latindex. The much larger Latindex coverage that was seen in Figure 6a is no longer present, which may indicate the predominance of Spanish and Portuguese journals in this bibliographic database.

By the investigation of only three established Latin American databases, it becomes evident that the local language scientific publishing is finding alternative outputs for publication. From this initial inspection, other databases might be explored, and a study by Alonso-Gamboa et al. (2012) highlighted some interesting alternatives for the future. Among them are LILACS, a Health Sciences database covering 901 journals from 26 countries: and PERIÓDICA, a database of around 1500 Science & Technology journals from Latin America and the Caribbean.

After that, there is still a pertinent question to answer: Are Brazilian researchers investigating significant research topics which are only covered by regionally relevant databases? To answer that, all titles from the 585.945 papers and reviews published in Brazil in 2013-2018 were exported to a corpus file which went through Neural Machine Translation into English. This corpus was used to generate a term map using VOSviewer software (van Eck et al., 2009), which can be seen in Figure 7.

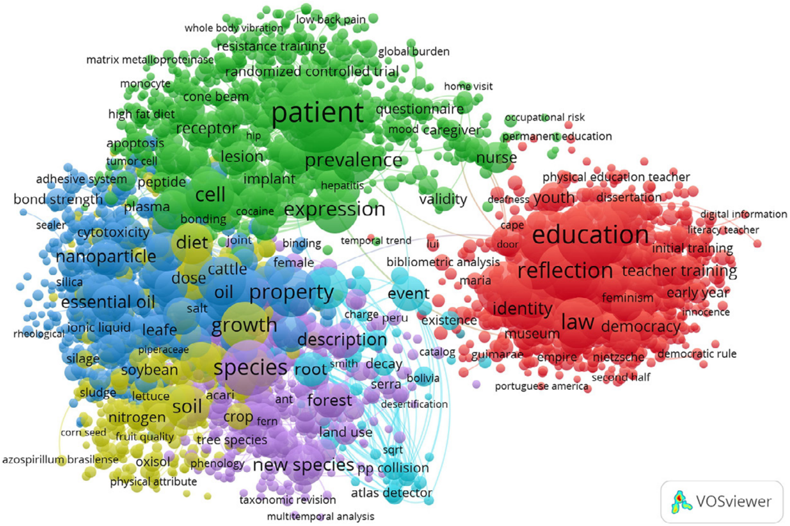


Figure 7. Term map generated from titles of Brazilian papers (2013-2018)

Figure 7 shows five different colour clusters with an evident relation to some of the main fields of science. Biomedical and Health Sciences appears in green and it is quite a significant cluster, in line with what has been seen about the Brazilian publishing profile. Other clusters seem to cover Life and Earth Sciences, with relevant topics related to Agricultural and Environmental Sciences, and a minor light-blue cluster goes into Physical Sciences and Engineering. Another large cluster is detached to the right of Figure 7, showing in red what evidently represents Social Sciences and Humanities. The fact that SSH research topics seem to be isolated from the rest becomes clearer with a colour overlay applied to the term map, as displayed in Figure 8.

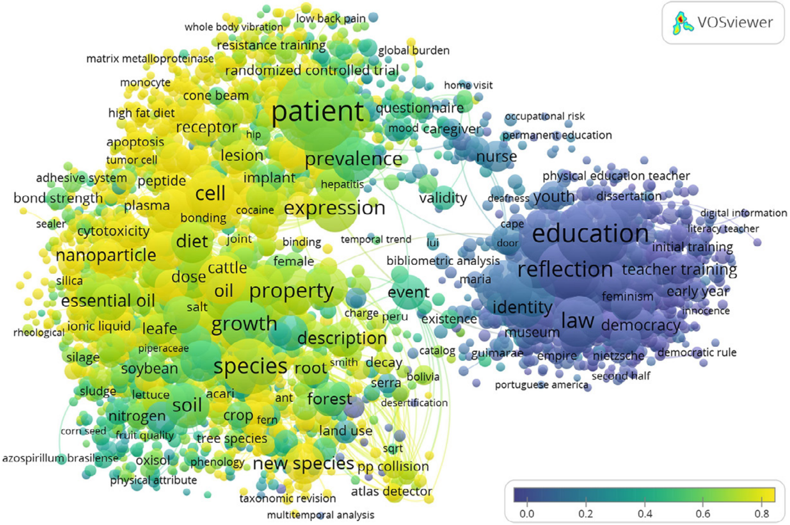


Figure 8. Term map generated from titles of Brazilian papers (2013-2018), colour-coded by Web of Science coverage

The overlay on Figure 8 was possible because, while exporting the publication titles to create the corpus, a score file was also produced, indicating if each individual paper was indexed by the Web of Science or not. From that, the SSH island seems even more detached, as most clusters are close to zero in the scale, meaning nearly all publications on the subject are not indexed by WoS. A small group of clusters in the left border of SSH reaches around 20% coverage, being weak links to other main fields of science.

Considering the left group of research topics displayed in Figure 8, some of the Biomedical, Health, and Life Sciences clusters shown at the top left, in light colours, are quite noticeable, as nearly 80% of the occurrence of such topics are in WoS indexed papers. Several other clusters are shown in shades of green, which mean their incidence is relatively balanced regarding WoS coverage, fluctuating between 40% to 60% in the presented scale.

Discussion and conclusions

A recent study conducted by Chavarro et al. (2018) investigated the extent to which the inclusion in the Web of Science would be an indication of journal quality. Among their findings was that journals with similar characteristics and editorial standards often receive unequal treatment due to their place of publication, discipline, and language. The authors warn research evaluators against the assumption that WoS and similar databases would assess journals without bias, joining several other scholars such as Alperín (2014), Garfield (1995) and Mounier (2018) in a call for caution in developing countries, with particular emphasis on Latin America.

In Brazil, policymakers and evaluators are engaged in ongoing discussions over a proposal for changes in the assessment of journals and publications that would trade a significant share of the qualitative methods currently adopted in favour of more journal-level indicators from WoS-like databases. This study adds to the warnings about qualitative dimensions of such databases and shows further science can be found in regionally relevant databases as well, where topics overlooked by the Web of Science are covered.

The data presented here show that WoS indexes around 41% of the qualified journals Brazilian researchers use to publish their work, accounting for just under 50% of the country's papers. Disciplinary distribution is hugely unbalanced, and the analysis of internal coverage for the broader WoS versus the Brazilian set of publications shows the existing biases in the database are taken to extremes in Brazil's case. That means disciplines such as the Social Sciences and Humanities, with already low coverage in the Web of Science, are even more underrepresented in Brazilian publications.

When it comes to using WoS indicators for evaluation, Moed and Visser (2007) describe four possible types of bibliometric studies from an internal coverage inquiry. The first type is known as a 'pure' WoS analysis, as it would be possible to rely solely on WoS source journals to analyse citation impact when internal coverage is over 80%. Types 2 and 3 – with coverage-ranges at 60–80 and 40–60, respectively – would require different levels of expansion in the sources covered, whether by including target articles not published in WoS source journals (2) or adding articles in proceedings volumes from a range of subsequent years (3). In case WoS coverage in a field is below 40, the mere value of a citation analysis based on WoS data should be questioned, even if target or source universes are expanded (Type 4 study).

The internal coverage analysis at the micro-field level classification for Brazil reveals that the bibliometric studies recommendations for the 4013 clusters would be: Type 1 – 43%; Type 2 – 29%; Type 3 – 14%; Type 4 – 14%. The problem is that of the 1741 clusters eligible for a 'pure' WoS analysis, 979 are in the Biomedical and Health Sciences, and 626 are in the Physical Sciences and Engineering, while only 19 are in Mathematics and Computer Science and eight in Social Sciences and Humanities. Even though for Mathematics, 48% of the clusters suggest a Type 2 study, 64% of those in SSH are ranked as Type 4. Consequently, we should question the mere value of the provided indicators for any quality assessment.

While these findings should be enough for policymakers designing Brazilian evaluation to step back from expanding their reliance on indicators from WoS-like databases, this research has also shown that most of the Brazilian scientific output not indexed by the Web of Science is written in Portuguese. That is particularly important for Brazil, as only a third of the researcher population states to have achieved some level of English fluency. Local language publications mean more people can produce and consume science, and societal impact may be even more relevant than citation metrics.

Since a significant share of local language papers could be found in Latindex, SciELO and RedALyC, future studies on identifying other high-coverage regional databases are recommended. Such databases cover various research subjects, some of them widely overlooked or just absent from the WoS indexed publications. Besides that, they usually have a core concern to promote the expansion of OA journals.

For instance, the Directory of Open Access Journals (DOAJ, 2021) – a community-curated online service with freely available data on OA details for 15.874 journals – reveals very interesting results when crossing its database with this project's data set. Around 62% of the Brazilian papers indexed by Latindex, SciELO or RedALyC were published in DOAJ listed OA journals. These are nearly 170.000 papers, of which almost 140.000 were in diamond OA, with no costs for authors. In contrast, only 25% of the set of articles indexed by WoS were published in DOAJ listed journals, and 63% of those were made open only through APC charges.

The presented numbers are consistent with a comprehensive study by Pavan and Barbosa (2018), which analysed OA publishing for Brazilian authored articles in WoS. Using data from the 2012-2016 period, the authors estimated 59% of OA papers in the database were in APC journals, at an average cost of USD 1492,27 per article. For a country like Brazil, where only 50% of PhD candidates receive monthly stipends (with values under USD 450), the high cost for OA publishing is a considerable problem. It's not that researchers would pay for publishing themselves, but the fact is the average APC cost of three papers could fund a whole year of stipends for a PhD candidate.

In conclusion, this research has shown that while adopting international indicators from established databases might seem like a good simple solution to improve local science, the reality is that databases such as WoS tell only half of the whole story. Brazilian scholarship is not only about state-of-the-art publications in top journals; it is also about regionally relevant topics, often destined to a Portuguese speaking audience; it is about access to publish and consume science, given the socioeconomical reality of the country. Brazilian science goes beyond the Web of Science, and a sound and comprehensive evaluation system should strive to capture its complexity, instead of trading it for restraining, short-sighted simplicity.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Alonso-Gamboa, J. O., & Russell, J. M. (2012). Latin American scholarly journal databases: a look back to the way forward. *Aslib Proceedings*.

Alperín, J. P. (2014). South America: Citation Databases Omit Local Journals, Nature, 511/7508: 155.

Barata, R. d. C. B. (2016). Dez coisas que você deveria saber sobre o Qualis. *Revista Brasileira de Pós-Graduação*, *13*(30), 13–40.

British Council. (2014). *Learning English in Brazil* (tech. rep.). British Council. São Paulo, Brazil. Retrieved February 12, 2021, from https://bit.ly/3b3lbeZ

CAPES. (2017). *Portaria nº 59, de 21 de março de 2017, Aprova o regulamento da Avaliação Quadrienal* (legislation). Diário Oficial da União.

CAPES. (2018). *Portaria nº 150, de 4 de julho de 2018. Institui o Grupo de Trabalho (GT) do Qualis Periódicos* (legislation). Diário Oficial da União.

CAPES. (2021a). *Dados Abertos* [DataSet]. Retrieved January 21, 2021, from https://dadosabertos.capes.gov.br

CAPES. (2021b). *Plataforma Sucupira*. Retrieved February 11, 2021, from http://sucupira.capes.gov.br/

Chavarro, D., Ràfols, I., & Tang, P. (2018). To what extent is inclusion in the Web of Science an indicator of journal 'quality'? *Research Evaluation*, *27*(2), 106–118.

DOAJ. (2021). *Public data dump* [DataSet]. Retrieved February 11, 2021, from https://doaj.org/docs/public-data-dump/

dos Santos, P. J. P., de Oliveira, T. M., de Araújo Jesus Paixão, F., Pascutti, P., Amado, A. M., da Hora Oliveira, D., …Mamiya, E. N. (2019). *Proposta do GT Qualis Periódicos*.

Education First. (2020). *EF English proficiency index: A ranking of 100 countries and regions by English skills* (tech. rep.). Retrieved February 12, 2021, from https://www.ef.nl/epi/

Garfield, E. (1995). Quantitative Analysis of the Scientific Literature and Its Implications for Science Policymaking in Latin America and the Caribbean, *Relation*, 29/1: 87–95.

Gibbs, W. W. (1995). Lost science in the third world. *Scientific American*, *273*(2), 92–99.

Hicks, D., Wouters, P. F., Waltman, L., de Rijcke, S., & Ràfols, I. (2015). The Leiden Manifesto for research metrics. *Nature News*, *520*(7548), 429–431.

Latindex. (2021). *Latindex Directory*. Retrieved January 21, 2021, from https://www.latindex.org/

Moed, H. F., & Visser, M. S. (2007). Developing bibliometric indicators of research performance in computer science: An exploratory study. *CWTS report*.

Mounier, P. (2018). 'Publication favela' or bibliodiversity? Open access publishing viewed from a European perspective. Learned Publishing, Learned Publishing, 31(S1), 299–305.

Nobre, L. N., & de Freitas, R. R. (2017). A evolução da pós-graduação no Brasil: histórico, políticas e avaliação. *Brazilian Journal of Production Engineering*, *3*(2), 18–30.

Pavan, C., & Barbosa, M. C. (2018). Article processing charge (APC) for publishing open access articles: the Brazilian scenario. *Scientometrics*, 117(2), 805–823.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Retrieved July 4, 2020, from https://www.R-project.org

RedALyC. (2021). *Sistema de Información Científica RedALyC: Colección de Revistas* [DataSet]. Retrieved February 3, 2021, from https://www.redalyc.org/coleccionHome.oa

SciELO. (2021). *SciELO Analytics* [DataSet]. Retrieved January 21, 2021, from https://analytics.scielo.org

Siqueira, M. B. (2019). Sucupira - a platform for the evaluation of graduate education in Brazil. *Procedia Computer Science, 146(2019)*, 247–255.

Sociedade Brasileira para o Progresso da Ciência & Academia Brasileira de Ciências. (2020). Qual o papel da CAPES na construção da ciência e da pesquisa no Brasil? [seminar].

Tijssen, R. J. W., Mouton, J., van Leeuwen, T. N., & Boshoff, N. (2006). How relevant are local scholarly journals in global science? A case study of South Africa. *Research Evaluation*, *15*(3), 163–174.

van Eck, N. J., & Waltman, L. (2009). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

Vasconcelos, S. M. R. de. (2008). *Ciência no Brasil: Uma Abordagem Cienciométrica e Lingüística* (Doctoral thesis). Rio de Janeiro: Universidade Federal do Rio de Janeiro.

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *arXiv preprint arXiv:2005.10732v2*.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392.

An approach to evaluate the success of individual research articles based on the Anna Karenina Principle

Zhen Cai[1] and Wen Lou[2]

[1] *kirsten_cai@163.com*
East China Normal University, Department of information management, Shanghai (China)
[2] *wlou@infor.ecnu.edu.cn*
East China Normal University, Department of information management, Shanghai (China)

Abstract

Evaluating the value of a research article is critical in scholarly communication. Recently the traditional methods of scientific evaluation based on informetric indicators and peer review have been challenged. Inspired by a theoretical method, entitled the Anna Karenina Principle (AKP), we propose an evaluation method to assess the research value by combining both qualitative and quantitative methods. In this RIP, we interviewed six researchers to testify the four dimensions of evaluating science studies, including the content value, the academic impact, the practical implication, and the society value. Combining with the AKP, we put forward a three-layers evaluation hierarchy for further quantitative measurement.

Introduction

Justifying the significance of research articles is a key process and an on-going activity in scholarly communication and scientific evaluation. At present, there are two common methods to evaluate individual research articles. One is informetric indicators, including citation, impact factors and other quantitative indicators. The other one is qualitative approach, which is represented by peer review. However, both of them have been undertaken with concerns in science community. The qualitative approach has the issue of the professional responsibility of reviewers and the expertise concerns (Grainger, 2007), which meets the challenge to unify the standard for giving opinions on article (Kangas & Hujala, 2015). Quantitative indicators cannot guarantee the quality of articles either (Bornmann & Daniel, 2008). Another unfortunate is that articles with potential breakthrough findings, which are most likely challenging the common sense, can be considered as the outlaw of scientific norms (Zeng et al., 2017). This phenomenon has happened in the history of science multiple times. Therefore, this paper aims to put forward a quantitative evaluation method based on a qualitative theory to evaluate the value and importance of individual research articles, and provides a new idea for improving the evaluation system of scientific research.

Research design

Theoretical foundation of evaluating individual articles

The essence of research articles evaluation is to evaluate the contents of the articles. The evaluation of individual research papers should undertake the level of semantic context as the core aided by various indicators from metadata level. Therefore, this paper puts forward an evaluation system of individual research articles from four dimensions.

(1) The content value of the article. The worth of a scientific paper should be reflected in the academic knowledge expounded in the article. The purpose of analyzing the content of the article is to find the valuable part except for the language of the article, then to go deep into the knowledge that the article attempts to achieve, and eventually to assess the potential academic and social value of this article (Evangelopoulos et al., 2012). Compared with other evaluation methods, it pays more attention to the comprehend the content of the article (Habib & Afzal, 2019) and can be applied to various sciences.

(2) The academic impact of the article. Currently, bibliometric indicators are the most commonly used method for research evaluation (Ellegaard & Wallin, 2015), which can be used to measure the academic impact of researches. One of the best ways to visualize the scientific development remains quantitative visualization. In such way, researchers normally could have a vision from (Lund, 2019). Even though the quantitative indicators are facing some challenges nowadays, the quantitative metrics can be a valuable assistant measurement for evaluation.

(3) The practical implications of the article. A valuable scientific paper not only needs to make novel discoveries, but also to give other researchers the opportunity to apply them. On the one hand, extensive application can give researchers more chances to discuss the article. On the other hand, it can actively promote the development of the discipline and provide favorable circumstances of the discipline (Prabhakaran et al., 2018).

(4) The society value of the article. Another key for research articles to be drawn upon attention by the society and public is that the authors present the work actively. With the rapid development of information technology and social media, some important researches will not only get attentions in the academic community, but also have great repercussions in the social community. In this way, researchers can speed up their careers when they building reputations in both areas (Fiala et al., 2017). In addition, research articles usually reflect somewhat the level of scientific output of the author's affiliations. And these organizations also have a certain impact on social development (Bana e Costa & Oliveira, 2012). Therefore, the evaluation of research articles cannot ignore its social value.

The philosophical relationship between the foundation and the AKP

Before proposing our evaluation system, a brief explanation of a theory involved in the system is needed. The Anna Karenina Principle (AKP) was introduced from science into scientific evaluation by Bornmann and Marx (2012) to describe that the success of a scientific work depends on many factors, and the absence of any one factor would lead to failure. Bornmann and Marx laid out eleven prerequisites: Solid evidence, Interest among colleagues who take up on the ideas, Verifying evidence from independent research groups, The paradigm should make possible (correct) predictions, Suitable techniques for the required measurements, A theory that provides a plausible explanation of the empirical findings, The paradigm is simple and elegant, The paradigm has great explanatory power, The paradigm has a catchy name, The last crucial step is achieved, and The researcher has stubbornness in thinking and good networking with colleagues in the field. According to this paper, if all these prerequisites mentioned above are included in a research, this research could to be considered to have made a great scientific outcome. If a research article can be considered with significant findings, this article must have met the prerequisites involved in the AKP.

These eleven prerequisites of the AKP are in line with the four values we proposed in this paper. First, the content value emphasizes that the evaluation of individual research articles needs to pay attention to the content of the article. The AKP offers a solid understanding of the evidence and the appropriate techniques, all of which can be gleaned from the context of articles. Secondly, AKP refers to the importance of citations in various prerequisites of AKP, which supports the academic impact of a study. Thirdly, the premise of the article implication is based on the reader's deep understanding of the article, therefore, the precise representation of an article is the key to present its implication value. The AKP addresses the importance of language and results presentation, which is corresponding to our third dimension. Finally, the AKP points out that a healthy community contributes to the success of scientific research in different infrastructure, which shows that scientific research needs the participation of both academics and society. Therefore, we adopt AKP to be a theoretical support for the evaluation method.

The structure of evaluation hierarchy

We conduct researcher investigations to decompose the four aspects of theoretical foundation we mentioned above and to investigate the consistency of AKP in the four aspects. In this primary research, we interviewed six subjects in different research areas of Library and Information Science, one in the US and other five in China. Each online interview was undertaken 30 to 40 minutes in December 2020. A total of eleven questions from the four aspects are as follows. The content and significance of the structural questions are illustrated.

(1) Understanding of the value of individual research articles. We would like to examine how researchers understand the most important value to a research article and the reasons in the theoretical sense. In this way, we can compare with our first aspect and AKP prerequisite, such as solid evidence.

- What are the factors you think the most important to conduct a research/a study/ to publish? For example: innovation, related research, research design, presentation of results, social impact, communication characterization, application potential.
- What would make you see a study as "Successful Science"?

(2) Understanding of the value of individual research articles through one's own scientific research experience. Unlike the first one, to testify the consistency with the implication value, this aspect focuses on either review practice as a reviewer or research process as a researcher.

- How do you think the current evaluation system in science community? What is your experience in terms of being a reviewer?
- Which section you pay the most attention on when you review a paper/a grant/an award? For example, originality, relevant research, research design, results representation, implication to the society, communication representation, usage potential.
- When you are required to evaluate a paper or a grant, which parts do you focus on most?
- How do you promote your own work? What are the factors that would influence your citing and promoting behavior?

(3) Understanding of the four dimensions in our evaluation system. After answering the above questions, we put forward the four dimensions in our evaluation system to these experts. The goal was to ask these interviewees whether the dimensions we proposed were a summary of their evaluation of individual research papers. If there are differences, we can also correct our hierarchy through these questions.

- How do you think the four dimensions?
- Expect for what you mentioned, do you think these four aspects can summarize your understanding?

(4) Finally, we inquired the subjects about the relationship between the four dimensions and the AKP. We aim to find out the resemblance between the four dimensions in our evaluation system and the AKP from others' perspective. Presumably, if we can confirm that there is indeed a philosophical relationship between the two, then the AKP can be regarded as an important theoretical support for the evaluation system proposed in this paper. Before asking the following questions, we introduced the basic information of the AKP to the subjects.

- How do you understand the AKP?
- Do you think the AKP can be a theoretical basis for evaluating individual research articles? Why and why not?

- What do you think is the relationship between the AKP and the four dimension we mentioned earlier?

We documented all audio files and converted them into texts. With the help of decomposing the interview text, we breakdown the four dimensions in line with the AKP into a preliminary evaluation system with three levels. Then we revisited three subjects to inquiry their opinions on the whole system and made further improvement.

Results

Through the understanding of the AKP and the interview with the LIS researchers, this paper comes up with a method with three levels which can be used to evaluate individual research papers. Level 1 is the basis of the hierarchy, which contains four dimensions. Both Level 2 and Level 3 are derived from the interviews with subjects and, based on discussions with the subjects, we corresponded these conditions to each of the four dimensions of Level 1. In this system, it can be found that from the four dimensions of Level 1 to the different requirements of Level 3, it is the result of further decomposition of the evaluation dimension by qualitative approach. The determination of Level 3 can also be used to prepare for further quantitative measurement. The evaluation hierarchy is shown in Figure 1.

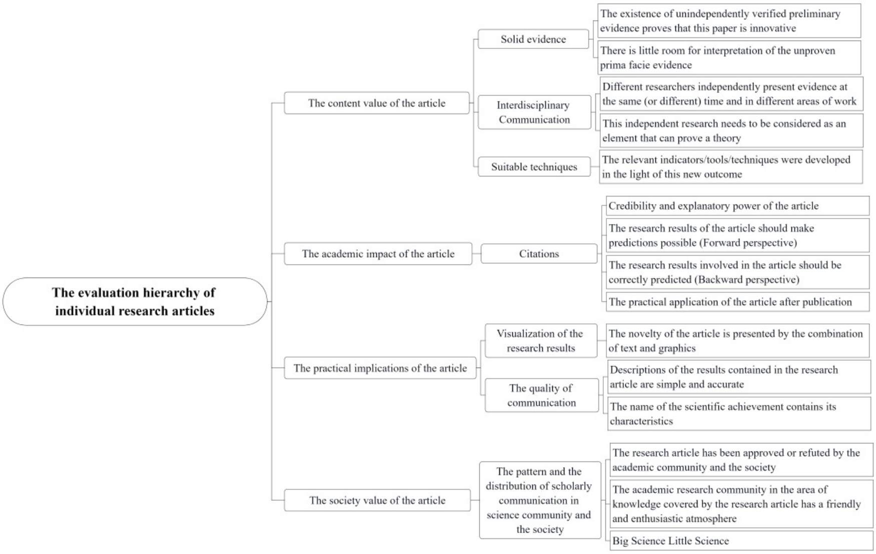

The evaluation hierarchy of individual research articles
The content value of the article
Solid evidence
The existence of unindependently verified preliminary evidence proves that this paper is innovative
There is little room for interpretation of the unproven prima facie evidence
Interdisciplinary Communication
Different researchers independently present evidence at the same (or different) time and in different areas of work
This independent research needs to be considered as an element that can prove a theory
Suitable techniques
The relevant indicators/tools/techniques were developed in the light of this new outcome
The academic impact of the article
Citations
Credibility and explanatory power of the article
The research results of the article should make predictions possible (Forward perspective)
The research results involved in the article should be correctly predicted (Backward perspective)
The practical application of the article after publication
The practical implications of the article
Visualization of the research results
The novelty of the article is presented by the combination of text and graphics
The quality of communication
Descriptions of the results contained in the research article are simple and accurate
The name of the scientific achievement contains its characteristics
The society value of the article
The pattern and the distribution of scholarly communication in science community and the society
The research article has been approved or refuted by the academic community and the society
The academic research community in the area of knowledge covered by the research article has a friendly and enthusiastic atmosphere
Big Science Little Science

Figure 1. The evaluation hierarchy of individual research articles

For the content value of the article, it includes three requirements in Level 2 and five requirements in Level 3. The first one is solid evidence. It is one of the prerequisites included in AKP, which means a scientific paper with significant research result must have adequate scientifically evidence to support the theory and/or method that the research is proposing. According to Bornmann (2012), core questions in science should be solved and steered in a new direction by the evidences which confirmed independently. In addition, novelty is the most common measurement to assess the value of a study, such scientific results should therefore have less room for interpretation in previous studies in the certain research areas. The interpretation of solid evidence also makes these two requirements a further step in this condition. The second requirement is interdisciplinary communication. Interdisciplinary communication refers to the specific research activities in which research subjects were created and interdisciplinary knowledge were developed according to the internal relations among disciplines. Currently, science community collaborate across disciplines. Even according to the Nobel Prize winners, there are no success in science if one only replies on one specific discipline. When evaluating the content of a research article, the reviewers may focus not only on how many researchers and theories this paper covers, but also on whether the theories proposed in this paper can be used as the theoretical basis for other disciplines. The last one is suitable techniques, which is also a vital prerequisite in AKP. A research could be expected to conduct new techniques for the measurement or verification. In response, the attention on whether new indicators or tools have appeared in the articles is important for evaluating.

For the academic impact of the article, citations are an important indicator of the quantitative analysis method, and citation indicators are increasingly applied in some subject areas to support qualitative evaluation (Kousha & Thelwall, 2011). For an article, the citation from different aspects can represent different meanings of the article. The structural interview also received several feedbacks on different applications of citation. We were told to conclude into four conditions. First is the credibility and explanatory power of the article, which can be presented in terms of the citation number of the article. The second is the analysis of the article from the aspect of forward perspective and backward perspective, which means whether the article has the potential to predict the trend of the subject through the citation distribution. Another one is that the practical application of the article after publication is related to the number of the citations of its application papers. Measuring these conditions can reflect the academic impact of individual scientific articles.

As for the practical implications of the article, we found that the subjects attach great importance to the presentation of the article and believe that good presentation can help promote the article. The visualization of research results and the quality of communication are the two factors affecting the application value of the article. If an obscure research article can be understood by other scholars or the general public, it shows that it has the possibility to be applied by others in some appropriate fields. It is easy for the readers to understand the content of the article by visual presentation of the combination of words and images, which has a positive effect on the further discussion and application of the article. A well promoted article must have composed of not only structured content but also a solid quality of communication skills. Additionally, a catchy name of the core idea, an easy-to-understand title, and a precisely simple description of research result would have been benefits for the article.

For the society value of the article, we focus on the pattern and the distribution of scholarly communication in both science community and the society. If a scientific article can be shared and discussed smoothly in both communities, it has the potential to promote not only the development of the discipline, but also the development of society. Meanwhile, it requires an enthusiastic and friendly atmosphere for the discipline to let researchers have opportunities of presenting articles in different places. Most importantly, the society value is reflected in the contribution to the society. If a scientific research is awarded as social significance, such as the Nobel Prize, the value must have been tested.

Conclusion and future research

In this research in progress, a new evaluation method was proposed by the understanding of the Anna Karenina Principle and a structural interview, which aimed to provide new ideas for the evaluation of individual research articles. We introduced a well-known theoretical approach of measuring success in sciences, the Anna Karenina Principle, as one of our fundamental theories. The prerequisites of AKP were examined to be in line with other theories in this paper. Combining with such in four dimensions, such as the content value, the academic impact, the practical implications, and the society value, we proposed a three-layers evaluation system, aiming to combine the qualitative analysis and quantitative analysis to improve the integrity and credibility of scientific evaluation.

The preliminary research will continue improving the hierarchy system by interviewing other researchers from outside of LIS area, government stakeholders, and public if necessary. The following research will focus on the quantitative measurement. Our current plan to decompose the Level 3 analysis is to continuously ask opinions on breakdown each requirement one by one with quantitative measures which we have data access or acquirable data. Sketchily, we will organize another round of interviews with questions, such as "How to determine the number of other disciplines that have discussed the research theory", "Can we transit the language property of the article into measurable indicators", "How to obtain the number of articles published in different approaches". We will gather all the actionable metrics or approaches. Last not the least, we will present the evaluation results as a way outside of box. Instead of giving weights to each indicator and resulting in another ranking or index, we will illustrate the results into the article context so that the readers would clearly have the image of the value and the impact of this article.

Acknowledgements

This research was supported by the Soft Science Key Project of Shanghai Science and Technology Innovation Action Plan 2020 (Grant No. 20692192500) and the open fund of ISTIC-CLARIVATE Joint Laboratory for Scientometrics.

References

Bana e Costa, C. A., & Oliveira, M. D. (2012). A multicriteria decision analysis model for faculty evaluation. *Omega, 40*(4), 424–436.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. In *Journal of Documentation* (Vol. 64, Issue 1).

Bornmann, L., & Marx, W. (2012). The Anna Karenina principle: A way of thinking about success in science. *Journal of the American Society for Information Science and Technology, 63*(10), 2037–2051.

Ellegaard, O., & Wallin, J. A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics, 105*(3), 1809–1831.

Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent semantic analysis: Five methodological recommendations. *European Journal of Information Systems, 21*(1), 70–86.

Fiala, J., Mareš, J. J., & Šesták, J. (2017). Reflections on how to evaluate the professional value of scientific papers and their corresponding citations. *Scientometrics, 112*(1), 697–709.

Grainger, D. W. (2007). Peer review as professional responsibility: A quality control system only as good as the participants. *Biomaterials, 28*(34), 5199–5203.

Habib, R., & Afzal, M. T. (2019). Sections-based bibliographic coupling for research paper recommendation. *Scientometrics, 119*(2), 643–656.

Kangas, A., & Hujala, T. (2015). Challenges in publishing: Producing, assuring and communicating quality. *Silva Fennica, 49*(4), 1–6.

Kousha, K., & Thelwall, M. (2011). Assessing the citation impact of book-based disciplines: The role of Google Books, Google Scholar and Scopus. *Proceedings of ISSI 2011 - 13th Conference of the International Society for Scientometrics and Informetrics, 1*, 361–372.

Lund, B. D. (2019). The citation impact of information behavior theories in scholarly literature. *Library and Information Science Research, 41*(4), 100981.

Prabhakaran, T., Lathabai, H. H., George, S., & Changat, M. (2018). Towards prediction of paradigm shifts from scientific literature. In *Scientometrics* (Vol. 117, Issue 3).

Zeng, C. J., Qi, E. P., Li, S. S., Stanley, H. E., & Ye, F. Y. (2017). Statistical characteristics of breakthrough discoveries in science using the metaphor of black and white swans. *Physica A: Statistical Mechanics and Its Applications, 487*, 40–46.

GeoAcademy: algorithm and platform for the automatic detection and location of geographic coordinates in scientific articles

Jesús Cascón-Katchadourian[1], Francisco Carranza-García[2], Carlos Rodríguez-Domínguez[3] and Daniel Torres-Salinas[4]

[1] *cascon@ugr.es*
University of Granada, Faculty of communication and documentation, Dept of Information and Comunication, 18071 Granada (España)

[2] *carranzafr@ugr.es*
University of Granada, ETSIIT, Dept of Software Engineering, 18014 Granada (España)

[3]*carlosrodriguez@ugr.es*
University of Granada, ETSIIT, Dept of Software Engineering, 18014 Granada (España)

[4] *torressalinas@ugr.es*
University of Granada, Faculty of communication and documentation, Dept of Information and Comunication, 18071 Granada (España)

Abstract

This document describes the GeoAcademy project, whose main objective is to automatically geolocate scientific articles, downloaded from general scientific databases such as Scopus or WoS. This geolocation is carried out on the content of the document, either through the capture of possible geographical coordinates that the document has, or toponyms that may appear in the document through an algorithm created for this purpose. In the methodology we explain the steps that have been taken in this project to create a sample database with articles that deal with Sierra Nevada (Spain) and the creation and design of the algorithm. The results show the technical data of the application of the algorithm on the database and its success rate, as well as a description of the platform created to graphically display the geolocated documents on a web map. Finally, in the discussion, we define the difficulties encountered, the possible bibliometric applications and its usefulness as a tool for viewing and retrieving information

Introduction

Geolocation is the possibility of locating information in specific geographic spaces to treat and process it (Ramos Vacca & Bucheli Guerrero, 2015). Technological advances in restitution by aerial photogrammetry and digitized cartography have contributed to this (Cortés-José, 2001), as have new procedures for cartographic writing and map printing, the appearance around 2005 of Google Maps, Bing Maps, OpenStreetMap and numerous programs to create maps such as OpenLayer, Leaflet, CartoDB or MapTiler (Cascón-Katchadourian & Ruiz Rodríguez, 2016). In the field of scientific evaluation and bibliometrics, geolocation techniques have been commonly used to locate scientific articles based on the origin of its authors (Catini et al., 2015). However, few studies have been devoted to analyzing a set of scientific publications based on their geographical coordinates and geolocating them on a map.

We would like to highlight here two current initiatives due to their importance and because they include several institutions behind their prototyping. One of them is GEOUP4, this "web portal shows on an interactive map the geolocation of the academic items of the repositories of the UPC, UPCT, UPM and UPV polytechnic universities grouped in the UP4 Association" (http://geo.up4. is/). The other project is JournalMap (https://www.journalmap.org/) a cooperative project between the USDA-ARS Jornada Experimental Range in Las Cruces, NM and the Idaho Chapter of The Nature Conservancy. This is one of the components of another tool called Landscape Toolbox. It is a project that geolocates scientific production based on the geographical location where the study is carried out to observe which areas have been studied and in which there are gaps. The present paper links with these two projects. Our study is

original and useful since it aims to geolocate the scientific articles in an automatic way through algorithms created for this purpose and not in a manual or semi-automatic way.
More specifically, there are three main objectives for this project, firstly (1) to develop an algorithm that allows the coordinates to be extracted from a collection of documents to identify exactly which places these studies deal with. The second objective (2) is that once the locations have been identified by the algorithm, they will be displayed on a map through an online platform. In the third objective (3), the platform will integrate a layer of bibliometric and/or altmetric information that will allow one to know the volume of production according to coordinates as well as different data on the scientific and social impact. It should be mentioned that in this paper will be presented only the results of objectives 1 and 2 applied to a set of scientific works that study the Sierra Nevada mountain range (Granada, Spain)

Material and methods

In order to create a collection of documentary information about Sierra Nevada, a search was done in the Scopus database. The Scopus database has been chosen for the facilities it offers for exporting full-text documents, thanks to Scopus Document Download Manager. With this type of search, we wanted to find scientific articles that dealt with the Sierra Nevada mountain range located in the province of Granada (Spain). As there were other mountain ranges with the same name in other parts of the world, the search for the term Spain or Granada was added. This search found 623 articles. The second step was the processing of the information for storage. After the pertinent manual verification work (the geolocation of the documents is automatic, not the selection of the sample), there are 447 documents with a complete associated pdf (not only the abstract or title) and which are openly accessible, the other 176 (623-447) records do not have an associated pdf, are incomplete or are not an open publication, of which 424 are about the Spanish Sierra Nevada and not about the Sierra Nevada of California or Peru. The third step was the identification of the coordinates. Although there are many types of coordinate systems, the most common that this project has found in the sample are the following types and subtypes of coordinates: the geographic coordinate system and the coordinate system UTM (Universal Transverse Mercator). The geographic coordinate system is subdivided into sexagesimal (degrees, minutes and seconds) and decimal (degrees and decimals). The UTM coordinate has multiple variants in that the authors express them in different ways, with or without the letters of the cardinal points, specifying the use or not (in our case it is 30S). Finally, in environmental studies, they usually use a subdivision of the use 30S, which is used in army maps and which is reflected in articles with VG, VF, WG and WF. Finally, our database contains 424 bibliographic references linked to their corresponding PDFs. This database has been used for the design and training of the algorithm.
For the geolocation of the contributions, we have developed a text mining algorithm based on the extraction of information through regular expressions and toponyms - keywords (see Figure 1). To carry out this automatic geolocation process, we design an algorithm based on 4 stages:

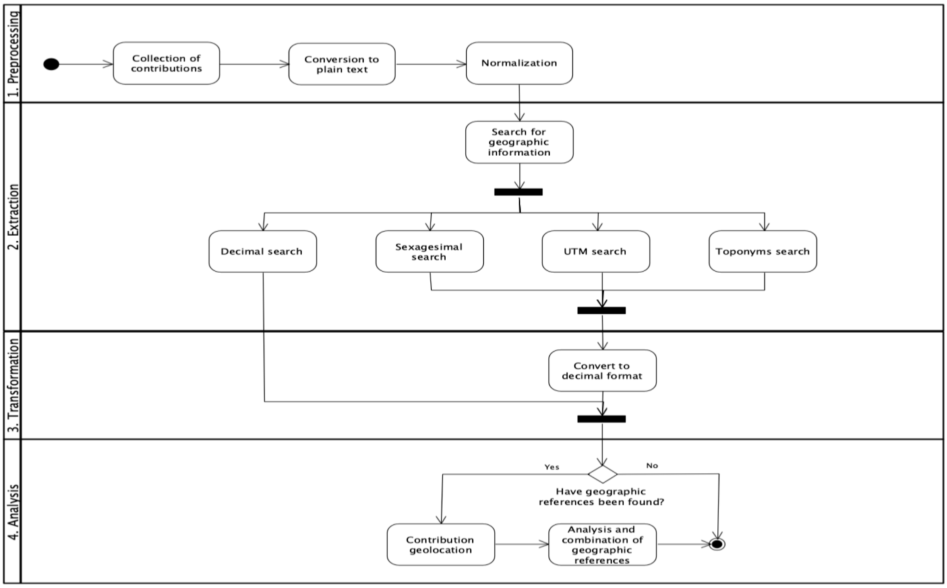
1. Preprocessing: In this stage, the contributions of the sample are processed in PDF format, converting the text of each one to a processable format (plain text), normalized and without all unnecessary information. For the conversion of PDF files to plain text, we have used pdftotext which is an open-source command-line utility.

2. Extraction: Search and extraction of geographical references in the text using two methods, (i) regular expressions for the search of geographical coordinates in their different formats (decimal, sexagesimal and UTM) and (ii) search of the frequency of appearance of place names. The database of toponyms for this study has been developed by combining all sources of public and georeferenced information of the Junta de Andalucía such as NGA and ITACA of the Andalusian Institute of Statistics and Cartography.

3. Transformation: In order to georeference the results obtained on an interactive map, all the results are processed to unify them in decimal format (latitude, longitude). To convert between the different geographic representations, we have used Proj4js1 which is a library to transform point coordinates from one coordinate system to another, including datum transformations.

4. Analysis: Finally, once we have all the information that has been extracted in the same format, a small analysis is carried out to decide, through heuristic behavior, if it is possible to geolocate the contribution. For this, we have mainly followed three criteria: (i) if geographic coordinates and place names have been found, we combine this information to select the coordinate that is textually closest to the most frequent place name; (ii) if only geographic coordinates have been identified, it is geolocated within the most referenced area; and (iii) if only toponyms have been recognized, we geolocate it in the one with the highest frequency as long as it exceeds a certain threshold (this threshold is calculated as the simple mean of the frequencies of the toponyms in the entire sample of this study)

Figure 1. Algorithm workflow.



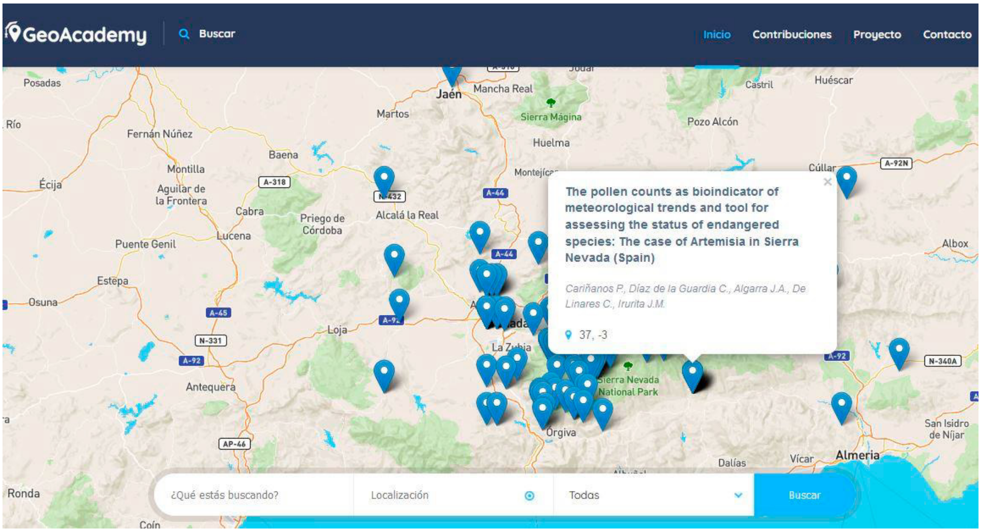
Results

Table 1 shows a summary of the results of the algorithm (figure 1) of automatic detection of geolocations being applied to the collection of 424 scientific articles on Sierra Nevada. In total, the algorithm has been able to geolocate 157 articles with coordinates, 37% of the total. One of the tools used to increase the number of geolocated articles and elements has been the use of place names that have allowed us to geolocate 5025 place names of 189 different articles. There are 2.6 place names on average per scientific article. Finally, the number of works that have been geolocated using the algorithm is 346, that is to say 81.6% of the original document collection has been located, either by coordinates or toponyms. It should be mentioned that one of the fundamental aspects to improve the success rate of the algorithm has been the use of place names.

Table 1. Indicators and general statistics in the algorithm training process

A) General indicators related to geolocation by coordinates	
A1. Number of articles analysed in the study:	424
A2. Number of articles containing geographical coordinates (157 + 43 (images) + 5	205
A.3 Number of articles containing geographical coordinates identified by the	157
A.4. Percentage of articles geolocated through geographical coordinates (A3/A1)	37.03%
B) Indicators related to geolocation by place names	
B.1. Number of place names identified by the algorithm	5025
B.2 Number of articles geolocated by place names	189
B3. . Percentage of articles geolocated by place names (B2/A1)	44.57%
C) Findings	
C.1 Number of geolocalisations achieved (coordenates + place names) (A.3+B.2)	346
C.2. Percentage of articles geolocated from the the total number of articles (C1/A1)	81.6%

Once objective 1 had been achieved, focus moved to the creation of an application to view scientific articles on a digital map based on the different coordinates they contain. The portal that has been developed currently contains the collection of documents on Sierra Nevada that has been analysed in this paper. The portal is operational at the following web address in beta format: https://geoacademy.everyware.es/.



The pollen counts as bioindicator of meteorological trends and tool for assessing the status of endangered species: The case of Artemisia in Sierra Nevada (Spain)

Cariñanos P., Díaz de la Guardia C., Algarra J.A., De Linares C., Irurita J.M.

37, -3

Figure 2. GeoAcademy application showing geolocated locations for the Sierra Nevada collection of articles.

As we can see in Figure 2, the geolocations detected by the algorithm are distributed, by clicking on each one we see the reference information of the paper and a link to view it. At the bottom you can filter by type of document you are looking for (article, conference, thesis, etc.) or you can search by keyword. The different tabs give us access to a traditional search engine with filters and a search radius, as well as a tab where the project and its members are described, as well as a contact form. Likewise, each scientific work that has been processed includes both the metadata of the databases (in this case Scopus) and all the metadata generated automatically by the algorithm. In our portal, the coordinates in decimal format as well as a complete list of the identified place names are added to the bibliographic description.

Discussion

This study contains a number of limitations that we proceed to list and which have mainly to do with processing. The first problem relates to the processing of PDF documents. We found several that did not have an associated pdf, for others the pdf requires payment, or only the title or the title and abstract are available. This limits the number of results you can cover. Secondly, the coordinates have some formal standardization, whereas in practice each field of knowledge has different guidelines for writing the coordinates, with greater variability with the UTMs that the algorithm has overcome without problems. Another problem is that several studies show the coordinates in image form, that makes them more difficult to extract, although in the future through OCR technology they could be captured. At the linguistic level, words with multiple means in the toponymy can be problematic, which can give both false positives and false negatives.

The success rate of the Geoacademy algorithm on all documents is 81.6%, a high percentage for this type of study. We will continue to train the algorithm with much larger documentary collections related to other geographical features. Likewise, it will be applied in other contexts, such as archaeological where most of the articles include precise geographic coordinates. It could also be applied to digital humanities, since interesting projects are being carried out where historical works are geolocated through the place names that appear in them.

The third objective of this work is to provide the platform with a layer of information capable of representing information of a bibliometric and altmetric nature, that is, of scientific and social impact. In this sense, future development implies providing the GeoAcademy platform with the following functionalities 1) Filter locations based on indicators (Impact Factor, Number of citations, Altmetric Attention Score ...) 2) Use of geoposition markers differentiated according to values of the indicators. Likewise, the platform will present a bibliometric summary of the different coordinates and locations, offering bibliometric information at a level not currently seen. Future developments could include its inclusion with an industry tool such as bibliometrix.

It would also be very interesting if users of large scientific databases such as Scopus or Web of Science, having performed a search with a list of results, could see said list geographically on a map, applying our algorithm for a better user experience. That is to say, to integrate our project in their platforms. Finally, numerous studies do not work with geographical points, we are currently working on how to show these studies on our platform

References

Cascón-Katchadourian, J., & Ruiz Rodríguez, A. Á. (2016). Descripción y valoración del software MapTiler: Del mapa escaneado a la capa interactiva publicada en la web.

Catini, R., Karamshuk, D., Penner, O., & Riccaboni, M. (2015). Identifying geographic clusters: A network analytic approach. Research policy, 44(9), 1749-1762.

Cortés-José, J. (2001). El documento cartográfico. En J. Jiménez Pelayo, J. Monteagudo López-Menchero, & F. J. Bonachera Cano. La documentación cartográfica : Tratamiento, gestión y uso (pp. 37-113). Huelva: Universidad de Huelva.

Ramos Vacca, I. D., & Bucheli Guerrero, V. A.. (2015). Automatic geolocation of the scientific knowledge: Geolocarti. Paper presented at the - 2015 10th Computing C*olombian Conference (10CCC),* pp. 416-424. doi:10.1109/ColumbianCC.2015.7333454

Star inventors: quantity and quality in the EOB model

Federico Caviggioli[1] and Boris Forthmann[2]

[1] *federico.caviggioli@polito.it*
Department of Management and Production Engineering (DIGEP), Politecnico di Torino -Corso Duca degli Abruzzi, 24, 10129, Torino, (Italy)

[2] *boris.forthmann@uni-muenster.de*
Institut für Psychologie in Bildung und Erziehung - WWU Münster - Fliednerstr. 21, D-48149 Münster (Germany)

Abstract

Star inventors generate superior innovation outcomes. Their capacity to invent high-quality patents might be decisive beyond mere productivity. However, the relationship between quantitative and qualitative dimensions has not been exhaustively investigated. The equal odds baseline (EOB) framework can explicitly model this relationship. This work combines a theoretical model for creative production with recent calls in the patentometrics literature for multifaceted measurement of the ability to create high-quality patents, extending the recent results obtained with forward citations. The results provide evidence in favor of the EOB across all quality indicators and that inventors' capacities to create quality patents can be measured to potentially identify star inventors in a way that explicitly takes the intricate relationship between overall productivity and patent quality into account. Furthermore, the rankings of inventors in terms of quantity or quality are overall more dissimilar than similar, confirming that quantity and quality can be measured as orthogonal dimensions. This result is of particular interest for organizations and for the society: incentives and compensation schemes that focus only on a quantitative assessment of inventors' output risk to neglect a relevant part of the innovation production.

Introduction

Star inventors are considered of extreme interest since they generate superior innovation outcomes (Groysberg & Lee, 2009; Oldroyd & Morris, 2012; Zucker & Darby, 1997). However, their identification is not immediate and can rely on different measures: is quantity of output sufficient or are quality indicators needed? Relatedly, disentangling inventor productivity and the quality of their patents is quite challenging because quantity and quality are intricately related (Forthmann et al, 2020a; Simonton, 2009). In addition, researchers have called for more multifaceted measurement of patent quality (Lanjouw & Schankerman, 2004; van Zeebroeck, 2011; Caviggioli et al., 2020).

In this study, we focus on patent inventors and their capacity to invent high-quality patents, which might be decisive beyond mere productivity (Rothaermel & Hess, 2007). The most frequent methods to define a star inventor in the literature considered when the inventor is either extremely prolific (quantity) or is involved in the creation of outstanding inventions (quality). However, the relationship between the two dimensions has not been exhaustively investigated in light of what can be considered determinant for the identification of stars as agents to increase the production of valuable innovations. The relationship between quantity and quality of inventions can be explicitly modeled within the framework of the equal odds baseline, EOB (Simonton, 2004, 1988). We use and extend the EOB framework to explicitly take the relationship between quantity and quality into account and thus provide useful diagnostic information for the identification of star inventors. Specifically, we incorporate latent variables into a multifaceted extension of the EOB that is based on multiple quality indicators derived from patent bibliometrics. This way the measurement of quality is isolated from quantity, implying that quantity and quality can be measured as orthogonal dimensions.

All the analyses are carried out at the individual level for a large sample of inventors with at least one US granted patent between 2008 and 2010. Different operationalization strategies are introduced to support the identification of prolific scientists and of outstanding inventors along different aspects of quality: technological complexity (through three indicators: the technological scope, the generality and originality indexes) and the value of an invention (measured by the number of citing families and by the geographical scope).

Research framework

Star individuals

The relevance of star scientists is not limited to a direct increase of output (Groysberg & Lee, 2009) but they also support organization activities (Kehoe & Tzabbar, 2015) and improve the attraction of resources and skilled personnel (Lacetera et al., 2004; Hess & Rothaermel, 2011). They also indirectly foster the productivity of peers and collaborators thanks to learning and emulation (Lockwood & Kunda, 1997). Although there is consensus on the presence of a general positive impact of star individuals, it is worth reminding that in some cases the literature identified negative effects in organizations due to coordination costs and conflicts (Groysberg et al., 2011; Swaab et al., 2014). Furthermore, hiring stars is often expensive (Groysberg et al., 2011) and thus it should be considered a critical activity. The findings of the literature support the need to improve the understanding of the way to identify exceptional scientists.

The identification of stars has taken different approaches in the literature, with respect to the examined field of activity and the different operationalizations of the criteria to distinguish outstanding from common individuals. In general, to be a star the individual must engage in disproportionately high performance relative to most other workers in their field (Aguinis & O'Boyle, 2014). The examined performance has been measured under different perspectives ranging from productivity (Lahiri et al., 2019; Zucker et al., 2002; Kehoe & Tzabbar, 2015), impact (Azoulay et al., 2010; Rothaermel & Hess, 2007) and, in some cases, visibility or celebrity (Oldroyd & Morris, 2012).

Star individuals have been studied in several contexts with particular attention to scientists/scholars (Azoulay et al., 2010) and inventors (Hohberger, 2016), thanks to data availability, namely articles and patents. Stars in these two categories have been similarly addressed by considering either their productivity in terms of quantity of output, in most cases through the number of articles or patents, their impact relying on a measure of quality such as the received citations (Liu, 2014; Hohberger, 2016; Hess & Rothaermel, 2011), or a combination of them (Kehoe & Tzabbar, 2015; Agrawal et al., 2017).

The equal odds baseline

The equal odds baseline (EOB) is a statistical model for the relationship between quantity and quality of scientific productions that is embedded within a comprehensive theoretical framework for scientific productivity (Simonton, 2009, 1988). In the EOB the number of inventors' high-quality patents H (i.e., the number of hits) is linearly regressed on the total number of patents T:

$$H = \rho T + u_i. \tag{1}$$

In Equation 1, ρ refers to the hit-ratio and u_i is a random error term for inventor i to take individual differences between inventors in hit-ratios into account (Simonton, 2009). The EOB further implies that individual hit-ratios H/T are uncorrelated with T because otherwise it would follow that the relationship between H and T is non-linear (Simonton, 2004). In other words, a positive linear correlation between H and T is a necessary but not sufficient condition for the EOB (Forthmann et al., 2020b, 2020a). The EOB further proposes an intercept of zero and a hit-ratio that equals the ratio of average H and average T. These implications of the EOB allow evaluation of model fit based on established criteria within the framework of structural equation modeling – *SEM* (Forthmann et al., 2020b, 2020a). The SEM approach to study the EOB has been further extended to allow quantifying if residual variance $Var(u)$ is larger as compared to a strict EOB with *Var*(*u*) = 0. This approach can be used to examine if individual differences are present in a given dataset (i.e., hit-ratio variance is larger than mere sampling error variation; Forthmann et al., 2020c). Indeed, the presence of individual differences in hit-ratios are a prerequisite to measure quality as a latent variable based on residuals resulting from multiple quality indicators.

A multifaceted extension of the equal odds baseline

Patents can be evaluated for a variety of quality criteria (Lanjouw & Schankerman, 2004; van Zeebroeck, 2011) and researchers have called for a multifaceted perspective on patent quality. However, the EOB has not yet been formulated in a way that takes multiple quality indicators into account. The current work aims at extending the EOB in this regard by formulating the EOB as a SEM in which individual differences in hit-rates are explained by a quality latent variable. In other words, the error term u_{ij} for inventor i (i = 1,…,I) and quality indicator j (j = 1,…,J) will be modeled by the following equation

$$u_{ij} = \lambda_j \eta_i + \varepsilon_{ij}, \tag{2}$$

with latent factor η_i as the capacity to create high-quality inventions, λ_j being the loading of the jth indicator on the capacity factor, and ε_{ij} the remaining error left unexplained after taking quantity and the capacity for quality into account. Inserting Equation 2 into Equation 1 yields a multifaceted EOB

$$H_{ij} = \rho_j T_i + \lambda_j \eta_i + \varepsilon_{ij}. \tag{3}$$

Specifically, the multifaceted EOB proposes that η and T are uncorrelated which allows for independent assessment of inventors' capacity for quality and productivity (i.e., quantity of output). As an extension of Equation 1, the model in Equation 3 can also be estimated within the SEM framework (Bollen, 1989). In SEM, a proposed path model and its implied covariance matrix and mean vector are examined for their discrepancy to their empirically observed counterparts. Useful models have a model-implied covariance matrix and mean vector that are close to the observed covariance matrix and vector of means (Bollen, 1989). Closeness of a proposed model and data in SEM can be evaluated by various established indices (West et al., 2012). In this approach, regression coefficients ρ_j and λ_j are estimated by maximum likelihood (or its robust variants), for example. Estimates of the latent capacity η_i can be obtained by means of empirical Bayes or Bartlett's method (Estabrook & Neale, 2013), for example. This model is illustrated for the five quality indicators used in this study (Figure 1 and further details in Section 3). In this illustrative example capacity for quality is unidimensional and if measurement of quality is assumed to support the identification of star inventors such a simple model might be appealing. However, the measurement of patent quality can be decomposed in two main dimensions: technological complexity and value (Caviggioli et al., 2020; van Zeebroeck & van Pottelsberghe, 2011). Technological complexity refers to the number of components, their degree of inter-dependence and decomposability (Singh, 1995; Wang et al., 2013). Technological complexity requires knowledge combinations from multiple domains (i.e., multidisciplinarity). Following this, the number of backward citations is often used as an indicator of technological complexity (van Zeebroeck, 2011; van Zeebroeck & van Pottelsberghe, 2011). Patent value as conceptualized in this work refers to the technical merit (forward citation count can be an indicator) of potential market size (as quantified by geographical scope) of the invention (e.g., Caviggioli et al., 2020a). Consequently, Equation 3 needs to be extended to a two-dimensional model that includes one latent variable to reflect technological complexity and a value latent variable. In the current work, we sought to empirically test the fit of inventor data to the EOB when quality is measured by multiple indicators and try to explain hit-ratio variation by means of a latent capacity to create high-quality inventions. Specifically, we wanted to compare a unidimensional model (i.e., quality capacity modeled as one latent variable; Lanjouw and Schankerman, 2004) with a two-dimensional model that reflects technological complexity and patent value (Caviggioli et al., 2020; van Zeebroeck & van Pottelsberghe, 2011).

Aim of the current research

The main aim of this study is to extend recent findings and theorizing on the EOB in several ways. First, we extend recent results obtained for forward citations of patents (Forthmann et al., 2020a) to other indicators because patent quality is a multifaceted construct (Lanjouw & Schankerman, 2004). Second, new EOB theorizing allows quantifying the amount of residual variance accounted by mere sampling variation (Forthmann et al., 2020c). This is useful to accurately estimate the amount of hit-ratio variation that is attributable to between-inventor differences. The presence of between-inventor variation in hit-ratios is essential for the measurement of capacity for patent quality. In this vein, the EOB is extended in this work to incorporate a latent quality variable that potentially explains individual differences in hit-ratios within a SEM framework. Finally, we aimed at comparing a unidimensional model with a two-dimensional model that incorporated latent variables to measure both technological complexity and patent value. Importantly, a reasonable fit of the data to either the unidimensional or two-dimensional multifaceted EOB implies that inventors' capacities to create quality patents can be measured to potentially identify star inventors in a way that explicitly takes the intricate relationship between overall productivity and patent quality into account.

Method

Data sources

The main data sources are PatentsView and PATSTAT. PatentsView is a data warehouse sourced from USPTO-provided data on published patent applications (2001-present) and granted patents (1976-present). It provides disambiguated inventors' names from the application of an algorithm that uses discriminative hierarchical coreference. Patent level data from PatentsView are linked to PATSTAT, the largest repository of patent data in terms of coverage and available information, maintained by the EPO with the collaboration of the main patent offices[1].

The analyses will be carried out at the level of inventors and the examined sample is defined by applying the following steps. The starting sample includes all the inventors with at least one US granted patent filed between 2008-10[2], corresponding to 725,577 disambiguated names in PatentsView. All the 4,297,710 granted patents associated to the selected inventors are linked to PATSTAT where further information is collected[3].

All the selected patents are associated to their INPADOC family (2.9 million families). Patent families represent a unit of analysis that is closer to invention: multiple patent documents regarding the same filing are collapsed to a single unit, providing a more accurate measure of inventors' productivity (OECD, 2009; Martínez, 2011). Furthermore, country extensions can be identified providing information on the geographical coverage. The earliest filing year and the IPC subclasses of each family are collected and several patentometrics are calculated following the approach described in (Caviggioli, et al., 2020). For each inventor it is thus possible to identify the portfolio of inventions and create portfolio level measures (these variables are described in the next section in detail) as of 2010 in terms of productivity. The cut-off year is required to consider a subsequent time window sufficiently large to calculate quality indicators such as citations and to account for potential delays in the publication of documents.

With the aim to clean the sample from potential errors in the original data, either in name disambiguation or in patent family identification, those inventors reporting a portfolio-level earliest filing date prior to 1981 (3.2%) were excluded. Inventors with no IPC codes associated to the portfolio were also eliminated (0.01%).

The final sample is a selection of 703,977 inventors active in the years 2008-10 and with a patenting history of maximum 30 years in 2010: each patent portfolios represent the cumulated inventions up to 2010. Table 1 reports the distribution of the portfolio size in the sample.

Table 1 Distribution of portfolios of inventions across selected inventors

Portfolio size [from	 to]	Number of inventors	Perc.	Cumulate
0	1	214226	30.4%	30.4%
2	5	252228	35.8%	66.3%
6	10	108368	15.4%	81.7%
11	50	119041	16.9%	98.6%
51	100	8032	1.1%	99.7%
101	500	2062	0.3%	100.0%
501	max	20	0.003%	100.0%
	Total	**703977**	**100%**	

Variables

Several measures of quality are calculated with the aim to test their relationship with the inventors' productivity (i.e. the number of patent families between 1981 and 2010). Each variable is computed at the level of patent family and in a second step is aggregated at the level of patent portfolio. Initially, we examined the fit of the data to the EOB for each of the five quality indicators in separation by means of correlations and a check of the presence of individual differences in the residual (Forthmann et al., 2020c). In a next step, we fitted the multifaceted EOB as a SEM model with the package lavaan (Rosseel, 2012) for the statistical software R (R Core Team, 2019). Model fit was based on fit indices such as the RMSEA, SRMR, CFI, and TLI according to existing cut-offs in the SEM literature (West et al., 2012). Using these fit indices is particularly helpful when examining the EOB in very large datasets because even small and negligible deviations from the EOB become easily statistically significant (Forthmann et al., 2020a). SEM model fit indices indicate if the data can be adequately described by the EOB when sample sizes are large. Finally, we estimated Raykov's reliability for the latent quality variables to quantify measurement precision (Raykov, 2009).

Table 2 provides an overview of the different variables that are introduced as indicators of various aspects of quality.

Technological complexity is described through three indicators: the technological scope, the generality and originality indexes. The technological scope counts the number of different IPC subclasses associated to patents (Lerner, 1994): the count is extended to the family level by considering all the family members. It provides a measure of multidisciplinarity: the broader the scope, the greater the complexity and the potential range of technological areas where it can impact (Harhoff et al., 2003). The originality and generality indexes were first introduced by (Trajtenberg et al., 1997) and are calculated considering the concentration of the different technological fields among the cited and citing patents of every focal document respectively. Both variables are computed in their unbiased version as described in (Hall & Trajtenberg, 2004) and IPC subclasses (four-digit IPC codes) are selected as technology fields. Coherently with the general approach, the patent citation network is generated at the level of INPADOC families, excluding intra-family citations. The generality index is a forward-looking measure describing the width of the technological advances. The originality index represents the scope of the underlying research.

The value of an invention is measured by the number of citing families and by the geographical scope. The forward citations provide a measure of the technical value (van Zeebroeck and van Pottelsberghe, 2011). The indicator considers only citations occurring in the first five years after the filing to account for the different time of exposure to the “risk” of receiving a citation (Caviggioli & Ughetto, 2016)[4]. The geographical scope indicates how large the expected market for the patented technology is. It is calculated as the number of jurisdictions in which patent protection is sought (Lanjouw et al., 1998; Agostini et al., 2015).

The construction of the indicators follows the approach proposed in van Zeebroeck (2011) but at the patent family level: each indicator of quality is calculated from the ranking of the examined patent family in a reference cohort, defined by a common technological sector and year (as a percentile). The reference sample consists of all the patent families associated to the US granted patents filed between 1981 and 2010. The sectors are identified by considering the concordance table between IPC codes and 35 technical fields developed by the WIPO[5]. The reference time is the earliest filing year among the family members. This approach allows calibrating the indicators with respect to each technological area and potential trends occurring in the time frame. When a patent family is associated to more technical fields, the indicator assumes the value of the highest percentile. For example, if an invention is developed in the fields “Optics” and “Pharmaceuticals” and the examined family ranks 60th for forward citations among all the inventions in the former and 80th in the latter, then the selected score for the considered family is 80.

Once all the indicators are calculated for each family, they are aggregated at the portfolio level. For each inventor, the number of outstanding inventions in the portfolio are counted. Such top inventions are identified when they are above the 95th percentile in the corresponding sector-year cohort (i.e. the family level indicators are equal or above 95). The number of excellent inventions according to each examined indicator is provided for all the inventors. Note that a single patent family could be above the excellence threshold in none, one or more of the indicators of quality.

Initially, we examined the fit of the data to the EOB for each of the five quality indicators in separation by means of correlations and a check of the presence of individual differences in the residual (Forthmann et al., 2020c). In a next step, we fitted the multifaceted EOB as a SEM model with the package lavaan (Rosseel, 2012) for the statistical software R (R Core Team, 2019). Model fit was based on fit indices such as the RMSEA, SRMR, CFI, and TLI according to existing cut-offs in the SEM literature (West et al., 2012). Using these fit indices is particularly helpful when examining the EOB in very large datasets because even small and negligible deviations from the EOB become easily statistically significant (Forthmann et al., 2020a). SEM model fit indices indicate if the data can be adequately described by the EOB when sample sizes are large. Finally, we estimated Raykov’s reliability for the latent quality variables to quantify measurement precision (Raykov, 2009).

Table 2 shows summary statistics of the variables. Additional variables are built to improve the model specifications and control for inventor’s characteristics. Since the selected sample includes inventors at different stage of their career a proxy of their expertise is introduced as the number of years since the first filing date. PatentsView database provides also data on inventors’ gender, as a result of the method explained in (Office of the Chief Economist, 2019). Note that data coverage is not complete (the gender is missing for 9.1% of the inventors in the examined sample).

Analytical approach

Initially, we examined the fit of the data to the EOB for each of the five quality indicators in separation by means of correlations and a check of the presence of individual differences in the residual (Forthmann et al., 2020c). In a next step, we fitted the multifaceted EOB as a SEM model with the package lavaan (Rosseel, 2012) for the statistical software R (R Core Team, 2019). Model fit was based on fit indices such as the RMSEA, SRMR, CFI, and TLI according to existing cut-offs in the SEM literature (West et al., 2012). Using these fit indices

is particularly helpful when examining the EOB in very large datasets because even small and negligible deviations from the EOB become easily statistically significant (Forthmann et al., 2020a). SEM model fit indices indicate if the data can be adequately described by the EOB when sample sizes are large. Finally, we estimated Raykov's reliability for the latent quality variables to quantify measurement precision (Raykov, 2009).

Table 2 Summary statistics of the examined variables

Measure	Indicator	Obs.	Mean	Std.Dev.	Min	Max
Quantity						
Inventor's patent portfolio size	Count of patent families having earliest filing year between 1981 and 2010	703,977	7.25	13.82	1	1041
Quality						
Excellent inventions as number of patent families above the 95th percentile in the portfolio in terms of:						
Technological complexity - Multidisciplinarity	Technological scope (count of IPC-4-digit)	703,977	0.26	0.87	0	103
	_ Share of portfolio		0.05	0.16	0	1
Technological complexity - Redeployability	Generality index	703,977	0.50	1.60	0	176
	_ Share of portfolio		0.07	0.20	0	1
Technological complexity - Novelty	Originality index	703,977	0.49	1.68	0	294
	_ Share of portfolio		0.07	0.20	0	1
Value – Technical merit	Count of Forward citations (5-years window)	703,977	0.59	1.93	0	198
	_ Share of portfolio		0.08	0.20	0	1
Value - Potential market size	Geographical scope (count of countries of extension in the family)	703,977	0.44	1.74	0	197
	_ Share of portfolio		0.07	0.20	0	1
Control variables						
Expertise – career	Years from the earliest filing date	703,977	8.00	6.76	1	30
Gender	Dummy=1 if male	640,043	0.87	0.34	0	1

Analysis

Univariate EOB examination

First, positive correlations between patent family counts (*T*) and all indicators of *H* were found (see column 1 in Table 3). This is in accordance with the EOB which proposes that the relationship between *H* and *T* is positive and linear. Expectedly, career length was moderately positively correlated with family count. In addition, career length correlated with all indicators of H. Correlations were small to moderate. Gender did not correlate with any of the creative productivity measures. Second, the correlation between patent family counts and all average quality indicators were close to zero (see column 1 in Table 4) which again provides evidence in favor of the EOB across all quality indicators. Importantly, both control variables correlated negligible small with all average quality indicators. As a final preparatory step prior to examination of the multifaceted EOB, we checked if residual variances were larger as expected under the strict EOB which implies constant hit-ratio. This check can only meaningfully be done when the EOB displays reasonable fit which was the case here based on correlations reported in Table 3 and Table 4. Residual variance findings are reported in Table 5. All observed residual variances (i.e., the variances of the u_i) were at least twice as large as compared to the minimum expected residual variance under the EOB (i.e., the residual variance under strict equal odds). Hence, we conclude that hit-ratio variation in the data was larger than expected under strict equal odds which only allows sampling error as a unique source of residual variation. Hence the data were promising for application of the multifaceted EOB with latent variable(s).

Table 3 Correlation matrix of examined variables: quality indicators expressed as number of excellent inventions in the portfolio (EOB relevant correlations in bold)

	Variables	1	2	3	4	5	6	7
1	Count of families	1.00						
2	Technological scope	**0.43**	1.00					
3	Generality index	**0.49**	0.65	1.00				
4	Originality index	**0.47**	0.61	0.78	1.00			
5	Count of fwd cit. (5-yrs)	**0.61**	0.47	0.46	0.42	1.00		
6	Geographical scope	**0.39**	0.33	0.25	0.22	0.41	1.00	
7	Career	0.38	0.24	0.29	0.26	0.24	0.18	1.00
8	Gender	0.05	0.03	0.04	0.03	0.03	0.02	0.11

Table 4 Correlation matrix of examined variables: quality indicators expressed as ratio of excellent inventions on total number of inventions in the portfolio (EOB relevant correlations in bold)

	Variables	1	2	3	4	5	6	7
1	Count of families	1.00						
2	Technological scope	**-0.03**	1.00					
3	Generality index	**-0.02**	0.33	1.00				
4	Originality index	**-0.02**	0.33	0.39	1.00			
5	Count of fwd cit. (5-yrs)	**0.00**	0.32	0.21	0.20	1.00		
6	Geographical scope	**-0.03**	0.23	0.09	0.09	0.19	1.00	
7	Career	0.38	0.04	0.08	0.07	0.08	0.03	1.00
8	Gender	0.05	-0.01	0.01	0.00	0.01	-0.01	0.11

Table 5 Estimates of the observed residual variance and the smallest expected residual variance.

Variables	Observed residual variance	Smallest residual variance
Technological scope	0.61	0.26
Generality index	1.94	0.46
Originality index	2.19	0.46
Count of fwd cit. (5-yrs)	2.35	0.54
Geographical scope	2.56	0.41

Multifaceted EOB results

The multifaceted EOB (without any latent quality variables) displayed excellent fit (RMSEA = .004, 90%-CI: [.004, .005]; SRMR = .007; CFI = .999; TLI = .999). Next, a unidimensional model with quality latent variable η was estimated and it displayed adequate fit (RMSEA = .015, 90%-CI: [.015, .015]; SRMR = .043; CFI = .960; TLI = .939). Reliability of the latent quality variable was .92 indicating excellent reliability. The latent quality variable η was predicted by career length to a small degree (β = .14, p < .001) and negligible small by gender (β = -.02, p < .001) with an overall R^2 = .02. As compared to the unidimensional model, the two-dimensional model with latent variables for value (η_{Value}) and technological complexity ($\eta_{\text{Technological complexity}}$) displayed better fit (RMSEA = .013, 90%-CI: [.013, .013]; SRMR = .034; CFI = .977; TLI = .962). The estimated standardized model parameters for this model are displayed in Figure 1. The value and technological complexity factors correlated with a value of .35. For the two-dimensional model, Raykov's reliability was .38 for the latent variable referring to value (see the loadings in Figure 1) and .96 for the latent variable referring to the technological complexity. Hence, value was not reliably measured, whereas technological complexity had excellent reliability. The technological complexity latent variable was predicted by career length to a small degree (β = .14, p < .001) and negligible small by gender (β = -.02, p < .001) with an overall R^2 = .02 mimicking the above findings for the unidimensional model. In addition, the R^2 for the value latent variable was zero which indicated that both control variables had a negligible relationship with value.

Identification of stars

We ranked the Top-10 inventors in Table 6 according to their productivity and the latent variable estimates based on the two-dimensional model. All inventors that occurred in more than one ranking list are depicted in bold font. Inventor ID221342 was the only one that occurred in all of the three lists in the Top-10 which highlights that stars may still be able to perform across a range of indicators even when some of them are orthogonal. Nonetheless, the rankings are overall more dissimilar than similar. Notably, the value latent variable estimates (empirical Bayes) were associated with the lowest measurement precision.

Table 6 Ranking of the Top-10 inventors based on productivity, general quality, value, and technological complexity.

Rank	Count of families (*T*)	η_{Value}	$\eta_{\text{Technological complexity}}$
1	ID193111	ID204114	***ID221342***
2	**ID136246**	ID239919	**ID57671**
3	ID55711	ID55136	**ID136246**
4	ID328897	ID19594	ID353689
5	**ID57671**	**ID72014**	ID44413
6	ID104847	***ID221342***	ID210006
7	ID12557	ID195909	ID153010
8	ID73327	ID52058	ID231139
9	**ID72014**	ID96931	ID96504
10	***ID221342***	ID18937	ID3808

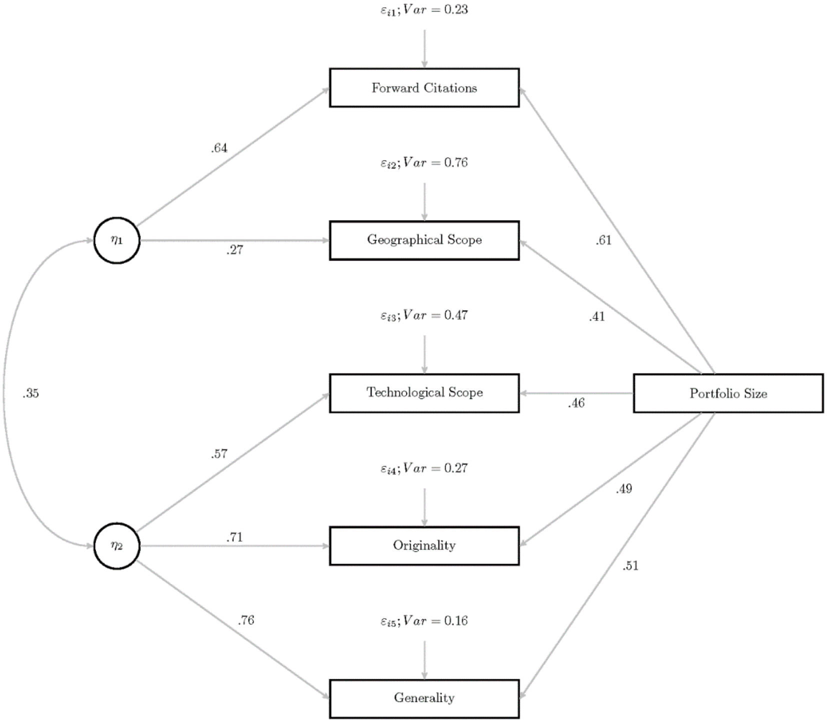

$\varepsilon_{i1}; Var = 0.23$
Forward Citations
.64
$\varepsilon_{i2}; Var = 0.76$
Geographical Scope
η_1
.27
.61
$\varepsilon_{i3}; Var = 0.47$
.41
.35
Technological Scope
.46
Portfolio Size
.57
$\varepsilon_{i4}; Var = 0.27$
.49
η_2
.71
Originality
.51
.76
$\varepsilon_{i5}; Var = 0.16$
Generality

Figure 1 Estimates for the two-dimensional path model of the multifaceted EOB (standardized parameters).

Conclusion

In this study, we employed the EOB framework to investigate the relationship between quantity and quality of inventions at the inventor level. This is driven by the relevance of star inventors as individuals able to generate not only more but also high-quality innovations. In particular, quality is decomposed along two main dimensions, technological complexity and value, by employing several patentometrics.

The results provide evidence in favor of the EOB across all quality indicators. Furthermore, the rankings of inventors in terms of quantity or quality are overall more dissimilar than similar, confirming that quantity and quality can be measured as orthogonal dimensions. This result is of particular interest for organizations and for the society. Incentives and compensation schemes that focus only on a quantitative assessment of inventors' output risk to neglect a relevant part of the innovation production which is more connected to the actual impact of the developed innovations.

Our work is not exempt from limitations and in particular the current empirical models do not account for organization level effects. The introduction of this further level of analysis would improve the results by considering the available resources. To perform such analyses, future studies will need to accurately identify patent assignees at the invention date and whenever inventors move to a different employer.

References

Agostini, L., Caviggioli, F., Filippini, R., & Nosella, A. (2015). Does patenting influence SME sales performance? A quantity and quality analysis of patents in Northern Italy. *European Journal of Innovation Management*, *18*(2). https://doi.org/10.1108/EJIM-07-2013-0071

Agrawal, A., McHale, J., & Oettl, A. (2017). How stars matter: Recruiting and peer effects in evolutionary biology. *Research Policy*, *46*(4), 853–867. https://doi.org/10.1016/j.respol.2017.02.007

Aguinis, H., & O'Boyle, E. (2014). Star performers in twenty-first century organizations. *Personnel Psychology*, *67*(2), 313–350. https://doi.org/10.1111/peps.12054

Azoulay, P., Zivin, J. S. G., & Wang, J. (2010). Superstar extinction. *Quarterly Journal of Economics*, *125*(2), 549–589. https://doi.org/10.1162/qjec.2010.125.2.549

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118619179

Caviggioli, F., & Ughetto, E. (2016). Buyers in the patent auction market: Opening the black box of patent acquisitions by non-practicing entities. *Technological Forecasting and Social Change*, *104*, 122–132. https://doi.org/10.1016/j.techfore.2015.11.031

Caviggioli, F., Colombelli, A., De Marco, A., & Paolucci, E. (2020). How venture capitalists evaluate young innovative company patent portfolios: empirical evidence from Europe. *International Journal of Entrepreneurial Behaviour and Research*, *26*(4), 695–721. https://doi.org/10.1108/IJEBR-10-2018-0692

Caviggioli, F., De Marco, A., Montobbio, F., & Ughetto, E. (2020). The licensing and selling of inventions by US universities. *Technological Forecasting and Social Change*, *159*, 120189. https://doi.org/10.1016/j.techfore.2020.120189

Estabrook, R., & Neale, M. (2013). A Comparison of Factor Score Estimation Methods in the Presence of Missing Data: Reliability and an Application to Nicotine Dependence. *Multivariate Behavioral Research*, *48*(1), 1–27. https://doi.org/10.1080/00273171.2012.730072

Forthmann, B., Leveling, M., Dong, Y., & Dumas, D. (2020a). Investigating the quantity–quality relationship in scientific creativity: an empirical examination of expected residual variance and the tilted funnel hypothesis. *Scientometrics*, *124*(3), 2497–2518. https://doi.org/10.1007/s11192-020-03571-w

Forthmann, B., Szardenings, C., & Dumas, D. (2020b). On the Conceptual Overlap between the Fluency Contamination Effect in Divergent Thinking Scores and the Chance View on Scientific Creativity: Fluency Contamination and Equal Odds. *The Journal of Creative Behavior*. Advance online publication. https://doi.org/10.1002/jocb.445

Forthmann, B., Szardenings, C., Dumas, D., & Feist, G. J. (2020). Strict Equal Odds: A Useful Reference to Study the Relationship between Quality and Quantity. *Creativity Research Journal*. Advance online publication. https://doi.org/10.1080/10400419.2020.1827605

Groysberg, B., & Lee, L. E. (2009). Hiring stars and their colleagues: Exploration and exploitation in professional service firms. *Organization Science*, *20*(4), 740–758. https://doi.org/10.1287/orsc.1090.0430

Groysberg, B., Polzer, J. T., & Elfenbein, H. A. (2011). Too many cooks spoil the broth: How high-status individuals decrease group effectiveness. *Organization Science*, *22*(3), 722–737. https://doi.org/10.1287/orsc.1100.0547

Hall, B. H., & Trajtenberg, M. (2004). *Uncovering GPTS with Patent Data*. NBER Working Paper Series.

Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition, and the value of patent rights. *Research Policy*, *32*, 1343–1363. https://doi.org/10.1016/S0048-7333(02)00124-5

Hess, A. M., & Rothaermel, F. T. (2011). When are assets complementary? Star scientists, strategic alliances, and innovation in the pharmaceutical industry. *Strategic Management Journal*, *32*(8), 895–909. https://doi.org/10.1002/smj.916

Hohberger, J. (2016). Does it pay to stand on the shoulders of giants? An analysis of the inventions of star inventors in the biotechnology sector. *Research Policy*, *45*(3), 682–698. https://doi.org/10.1016/j.respol.2015.12.003

Kehoe, R. R., & Tzabbar, D. (2015). Lighting the way or stealing the shine? An examination of the duality in star scientists' effects on firm innovative performance. *Strategic Management Journal*, *36*(5), 709–727. https://doi.org/10.1002/smj.2240

Lacetera, N., Cockburn, I. M., & Henderson, R. (2004). Do Firms Change Capabilities By Hiring New People? a Study of the Adoption of Science-Based Drug Discovery. *Advances in Strategic Management*, *21*, 133–159. https://doi.org/10.1016/S0742-3322(04)21005-1

Lahiri, A., Pahnke, E. C., Howard, M. D., & Boeker, W. (2019). Collaboration and informal hierarchy in innovation teams: Product introductions in entrepreneurial ventures. *Strategic Entrepreneurship Journal*, *13*(3), 326–358. https://doi.org/10.1002/sej.1331

Lanjouw, J. O., & Schankerman, M. (2004). Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators*. *The Economic Journal*, *114*(495), 441–465. https://doi.org/10.1111/j.1468-0297.2004.00216.x

Lanjouw, J. O., Pakes, A., & Putnam, J. (1998). How to count patents and value intellectual property: The uses of patent renewal and application data. *Journal of Industrial Economics*, *46*(4), 405–432. https://doi.org/10.1111/1467-6451.00081

Lerner, J. (1994). The Importance of Patent Scope: An Empirical Analysis. *The RAND Journal of Economics*, *25*(2), 319. https://doi.org/10.2307/2555833

Liu, K. (2014). Human Capital, Social Collaboration, and Patent Renewal Within U.S. Pharmaceutical Firms. *Journal of Management*, *40*(2), 616–636. https://doi.org/10.1177/0149206313511117

Lockwood, P., & Kunda, Z. (1997). Superstars and Me: Predicting the Impact of Role Models on the Self. *Journal of Personality and Social Psychology*, *73*(1), 91–103. https://doi.org/10.1037/0022-3514.73.1.91

Martínez, C. (2011). Patent families: When do different definitions really matter? *Scientometrics*, *86*(1), 39–63. https://doi.org/10.1007/s11192-010-0251-3

OECD (2009). OECD Patent Statistics Manual. In *OECD Patent Statistics Manual*. https://doi.org/10.1787/9789264056442-en

Office of the Chief Economist. (2019). Progress and Potential: A profile of women inventors on U.S. patents. *IP Data Highlights*, *2*, 84–86.

Oldroyd, J. B., & Morris, S. S. (2012). Catching falling stars: A human resource response to social capital's detrimental effect of information overload on star employees. *Academy of Management Review*, *37*(3), 396–418. https://doi.org/10.5465/amr.2010.0403

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Raykov, T. (2009). Evaluation of Scale Reliability for Unidimensional Measures Using Latent Variable Modeling. *Measurement and Evaluation in Counseling and Development*, *42*(3), 223–232. https://doi.org/10.1177/0748175609344096

Rosseel, Y. (2012). lavaan : An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Rothaermel, F. T., & Hess, A. M. (2007). Building dynamic capabilities: Innovation driven by individual-, firm-, and network-level effects. *Organization Science*, *18*(6), 898–921. https://doi.org/10.1287/orsc.1070.0291

Simonton, D. K. (2004). *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. Cambridge University Press.

Simonton, D. K. (2009). Scientific Creativity as a Combinatorial Process: The Chance Baseline. In P. Meusburger, J. Funke, & E. Wunder (Eds.), *Milieus of Creativity* (Vol. 2, pp. 39–51). Springer Netherlands.

Simonton, D. K. (1988). *Scientific genius: a psychology of science*. Cambridge University Press.

Singh, K. (1995). The Impact Of Technological Complexity And Interfirm Cooperation On Business Survival. *Academy of Management Proceedings*, *1995*(1), 67–71. https://doi.org/10.5465/ambpp.1995.17536285

Swaab, R. I., Schaerer, M., Anicich, E. M., Ronay, R., & Galinsky, A. D. (2014). The Too-Much-Talent Effect: Team Interdependence Determines When More Talent Is Too Much or Not Enough. *Psychological Science*, *25*(8), 1581–1591. https://doi.org/10.1177/0956797614537280

Trajtenberg, M., Henderson, R., & Jaffe, A. (1997). University Versus Corporate Patents: A Window On The Basicness Of Invention. *Economics of Innovation and New Technology*, *5*(1), 19–50. https://doi.org/10.1080/10438599700000006

van Zeebroeck, N. (2011). The puzzle of patent value indicators. *Economics of Innovation and New Technology*, *20*(1), 33–62. https://doi.org/10.1080/10438590903038256

van Zeebroeck, N., & van Pottelsberghe de la Potterie, B. (2011). The vulnerability of patent value determinants. *Economics of Innovation and New Technology*, *20*(3), 283–308. https://doi.org/10.1080/10438591003668638

Wang, Y., Zhou, Z., & Li-Ying, J. (2013). The impact of licensed-knowledge attributes on the innovation performance of licensee firms: evidence from the Chinese electronic industry. *The Journal of Technology Transfer*, *38*(5), 699–715. https://doi.org/10.1007/s10961-012-9260-0

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). The Guilford Press.

Zucker, L. G., & Darby, M. R. (1997). Individual action and the demand for institutions: Star scientists and institutional transformation. *American Behavioral Scientist*, *40*(4), 502–513. https://doi.org/10.1177/0002764297040004012

Zucker, L. G., Darby, M. R., & Armstrong, J. S. (2002). Commercializing knowledge: University science, knowledge capture, and firm performance in biotechnology. *Management Science*, *48*(1), 138–153. https://doi.org/10.1287/mnsc.48.1.138.14274

[1] More information at www.patentsview.org (last access and data release in August 2020, disambiguated inventors' names updated in March 2020) ad at https://www.epo.org/searching-for-patents/business/patstat.html (last access in August 2020, version of database used in this study: fall 2019).

[2] Only "utility" patents have been considered. Withdrawn patents are included (corresponding to 0.17% of the examined granted patents).

[3] In terms of granted patents, the selected sample is 58% of the total US granted patents in PatentsView (1975-2020)

[4] The models were also tested considering a variable with a 10 years window to capture citations (as in Forthmann et al., 2020a). The results are very similar and are available on request. Note that intra-family citations are not considered.

[5] Source: WIPO IPC-Technology Concordance Table (last update in 2016).

Connecting Brain and Heart: Artificial Intelligence for Sustainable Development

Diego Chavarro[1], Jaime Andrés Pérez-Taborda[2] and Alba Ávila[3]

[1]*dchavarro@gmail.com*
Colombian Society of Engineering Physics (SCIF), 660003, Pereira (Colombia)

[2]*ja.perezt@uniandes.edu.co*
Colombian Society of Engineering Physics (SCIF), 660003, Pereira (Colombia)
Microelectronics Center (CMUA), Department of Electrical and Electronic Engineering, Universidad de Los Andes, 111711, Bogota (Colombia)

[3]*a-avila@uniandes.edu.co*
Microelectronics Center (CMUA), Department of Electrical and Electronic Engineering, Universidad de Los Andes, 111711, Bogota (Colombia)
Colombian Society of Engineering Physics (SCIF), 660003, Pereira (Colombia)
(Corresponding author)

Abstract

A key objective of global policies on Artificial Intelligence (AI) is to foster AI research for sustainable development (SD). In this paper we examine the inclusion of this objective in AI published research by addressing three questions: 1) To what extent is AI research addressing the sustainable development goals (SDGs)? 2) Which subject areas of AI show an emerging interest in SD? And 3) What patterns of collaboration between regions of the world are being stimulated by AI? We analyzed AI papers from 2000 to 2019 using the IEEE Xplore database by: 1) Identifying the number of AI papers that address SDGs in their titles, abstracts, and keywords. 2) Developing a composite indicator based on the number of documents produced, citation impact, and inventive impact to distinguish areas with an emerging interest in SD; 3) Exploring co-authorships at three levels: region, income group, and country. The results show that a small share of papers is explicitly focused on SD, but there is an emerging interest in *Broadcast technology, Systems Engineering and Theory, Ultrasonics, Ferroelectrics, and Frequency Control, Sensors,* and *Education*. Inter-regional and inter-income group collaboration are limited, and network power is concentrated in a few countries. The results could be useful to align the connection between technical knowledge, strategic planning for S&T investment, and SD policies.

Introduction

The fourth industrial revolution is characterized by the exponential development of technologies to radically transform life on Earth (Schwab 2017). This new revolution has been advocated as an opportunity to improve business productivity, automate various manual trades, provide greater control over the quality of production, among others (Mou 2019). On the other hand, there is growing concern about the adverse effects of this revolution on the environment, the economy, and society (Harari 2017). The transformations that the fourth industrial revolution is having in the world make it important to understand its risks and potential benefits, and the role of academia, industry, government, and civil society in steering it.

From a global perspective, there is a competition for leadership in the fourth industrial revolution's determining technologies. For example, most OECD member countries are improving their digital infrastructure, educating their citizens in information technologies, supporting new technology-based companies, and research in "convergent" and "exponential" technologies (Planes-Satorra & Paunov 2019). The confrontation between China and the United States over the development and control of 5G technology (Sørensen et al. 2016) is an extreme but revealing example of the global competition situation. Consequently, several national and international agencies have begun to carry out policy studies to understand how to insert themselves into the new global value chain that is shaping the fourth industrial revolution and what international cooperation strategies can be more effective to position themselves as leaders in the technologies that drive this revolution (Kagermann et al. 2016; UNIDO 2018; Navarro 2018). Although this quest for leading technological development is valid, they tend to focus on the technical aspects of it, leaving aside other important concerns.

One such concern is if technological breakthroughs will work towards planetary well-being. If so, it is imperative to reconcile technological advancement with societal purposes (Cassi et al. 2017). Researchers have an essential role in achieving this reconciliation because they produce a good part of the enabling knowledge behind technology, and it is not uncommon that they produce the technologies themselves (Breschi et al. 2005; Zellmer-Bruhn et al. 2021).

Technologies are often seen as mere "technical" developments, but they also involve social aspects that need to be reflected upon by engineers (European Parliament 2020). For instance, AI algorithms to detect potential criminals can be biased by human assumptions on the physical appearance of people or their gender (Osoba & Welser 2017). Therefore, reflecting beyond the technical aspects of technology is needed to produce humanistic technologies. This demands engineers to think of technologies to reduce our impact on the planet because, as Jane Goodall puts it, "only when our clever brain and our human heart work together in harmony can we achieve our true potential" (Goodall 2014).

Part of achieving the true potential that Goodall talks about implies producing sustainable technologies with a clear purpose to ensure sustainable development. Therefore, it is crucial to understand to what extent engineering research, which is behind the significant breakthroughs in world technology, addresses sustainable development. In this paper, we understand sustainable development as development that aims to achieve a social, environmental, and economic balance (Chavarro et al. 2017).

We focus on AI as a ground-breaking domain of technologies and research with the power to redefine global society. Recently, the OECD issued a crucial policy paper in which they recognize that "AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being" (OECD 2020). The IEEE has also made available a clear position to "Support the R&D needed to advance innovation and development in artificial intelligence and its application to benefit humanity" (IEEE 2019). Our work is a contribution to a broad question, which remains unexplored, on the extent to which "the digital era could contribute to or jeopardize sustainable landscapes (…), what is the role of digital technologies in pursuing the (2030) Agenda and how to enable a sustainable digitalization" (del Río Castro et al. 2020). Within this context we explore the following questions:

1. To what extent is sustainable development addressed in AI research produced globally?
2. Which subject areas of AI show an emerging interest in sustainable development?
3. To what extent does AI stimulate collaboration between regions of the world?

We adopt a "bottom-up" approach by identify what is already happening to achieve both technological advancement and planetary health. We look at the publications on AI available through the Institute of Electrical and Electronics Engineers - IEEE Xplore. Investigating AI from this perspective may help formulate evidence-based public policies to foster the developments that are already taking place, better understand the disciplinary map that is being configured in the world of AI, and explore the extent to which the intellectual work of engineering researchers is considering the wider uses of their technologies and their impacts on the environment and society.

Methodology

This paper focuses on AI as an enabling technology (European Patent Office 2020) that will experience exceptional growth over the next years (Klavans, Boyack, & Murdick 2020). This specific field is also one of the subjects for which there will be sustained policy support (European Commission 2020).

There are different definitions of AI (see Nilson 2010; IEEE 2019). Given this plurality, we do not attempt a conceptual definition of it. However, we acknowledge that there are at least four domains of AI that cover the range of inventions and research on the subject. These are: acting like a human, thinking like a human, acting rationally, and thinking rationally (Russell, 2010, p. 5). This understanding of AI shows its complexity, which involves multiple definitions for a range of AI applications (i.e. thinking like humans could be related to algorithms and techniques for the development of chat bots, acting like humans the humanoid robots, and autonomous robots). Efforts to develop standard AI definitions in different technological niches and social domains can be seen in IEEE (2021a).

Despite the broad spectrum of areas covered by AI as a socio-technical phenomenon, part of AI research happens within disciplines in the Engineering domain, which has established authorities, ontologies, and classifications. Because of its usefulness to locate AI within the Engineering knowledge domain, in this work we consider AI from the IEEE thesaurus perspective, which is an authoritative source for different disciplines of Engineering, including AI (IEEE 2021a). We, then, identify AI papers as those classified within the field of Artificial Intelligence by the IEEE thesaurus (IEEE 2021a).

Sample

Table 1. Number of documents on AI per year

Publication year	Number of documents	Publication year	Number of documents
2020	27,783	2009	8,606
2019	39,269	2008	7,585
2018	25,800	2007	6,022
2017	17,503	2006	4,961
2016	11,000	2005	4,552
2015	9,797	2004	3,539
2014	8,736	2003	2,740
2013	8,064	2002	2,458
2012	8,747	2001	1,805
2011	9,591	2000	2,737
2010	8,852		

Our sample consists of papers developed in the Engineering sciences produced from 2000 to 2020 and indexed by the IEEE Xplore database with at least one of the subjects being Artificial Intelligence (AI) (IEEE 2021a). Our database choice has to do with the availability of fine-grained classification of papers within AI that this database provides, which is required to produce meaningful analysis for experts in the discipline. Other more general citation databases could offer better interdisciplinary coverage, but the subject classifications may belong to non-Engineering perspectives on AI. In total, we have acquired 220,147 records of research articles, reviews, proceedings, and books available in this database. Table

1 shows the number of documents gathered per year and Table 2 the number of documents by type of publication[1]:

Table 2. list of documents by type

Content-Type	Number of Documents
Conference proceeding	174,694
Journal papers	40,509
Magazine papers	2,959
Early access articles	1,291
Books	634
Standards	38
Courses	20

Data classification

Two types of classifications were fundamental to understand the relationship between sustainable development and AI in engineering. The first type is a subject classification, for which the IEEE Thesaurus was used. IEEE has also made available a taxonomy aimed explicitly at describing engineering subjects. The IEEE taxonomy is derived from the IEEE Thesaurus, which is a controlled vocabulary of 10,644 terms. The taxonomy provides a hierarchical classification that represents the relationship between terms. As expert committees develop the thesaurus and the taxonomy, they integrate expert knowledge and present data in a meaningful context to engineering research communities. Although this paper is about Artificial Intelligence, we used the whole IEEE Taxonomy with its 52 categories (see IEEE 2021b) because AI papers apply different engineering knowledge fields. For each paper, we identified its root term, first, second, and third levels according to the taxonomy.

The second classification is less standard because it is aimed at identifying papers that address sustainable development. Sustainable development is not a disciplinary field, but a domain of interest linked to a long-term policy agenda, with a clear direction towards a better balance between human existence and the environment, economy, and society (Chavarro et al. 2017). For this reason, there is no clear standard or agreement on what knowledge contributes to sustainable development. In principle, all knowledge could contribute to sustainable development if it is used to improve that balance. However, some knowledge shows this intent more explicitly, for instance, in a paper that attempts to build a clean alternative to energy production. In contrast, some knowledge may not contribute directly or may even hinder sustainable development, as in the case of knowledge to produce mass destruction weapons.

Even though the classification of papers according to their relationship to sustainable development remains a challenge (Armitage & Mikki 2020), two main approaches have been developed by researchers for this task. The first one is through keyword searches curated by experts in the field (Elsevier 2020). This approach offers textual identification of elements related to sustainable development, but not semantic identification of concepts. A second approach uses machine learning algorithms to classify papers based on their similarity to a set of papers classified by humans (OSDG 2021). This approach offers a conceptual identification of papers. Still, the models can become inaccurate because of ambiguity,

[1] The search was performed over all publication fields in the database, using "artificial intelligence". The data was obtained through the IEEE Xplore API using PhP.

different sizes of the categories, and the difficulty in understanding what the algorithms' criteria favor given the entered data.

As both models have pros and cons, we performed the two classification approaches. For the keyword classification, we used Elsevier's keywords, and for the machine learning approach, we used OSDG's (OSDG 2021), which is an initiative to provide a service for text classification into the categories of the sustainable development goals. OSDG approach uses Microsoft Academic Knowledge Project and other resources as a means for topic classification. In total, the keyword approach produced 44,563 papers compared to 84,309 produced by the machine learning approach. We examined the results of the two approaches and found that the keyword approach offered classifications more consistent with our understanding of sustainable development in a random set of 1,000 papers checked manually by the team for this dataset. For this reason, we use the keyword approach here and kept the machine learning approach for further research.

In addition to discipline subject and sustainable development, we classified the papers according to different geographical categories: countries, regions, and income groups. For this, the World Bank's classification of countries was used. As the IEEE database's countries are not normalized, it was necessary to identify and extract country names from the affiliations field. This was done in two ways: first, identifying mentions of country names and well-known cities in the author affiliations field. The remaining papers for which there was an affiliation field were identified by leveraging the first classification's knowledge base, querying full-text indices to return similar records to the ones already classified. This approach provided an accuracy of classification of 97% in a random sample of 1,000 records checked manually. 17,555 out of 220,147 articles did not have an affiliation field, or it was not possible to identify their countries.

Analysis

To explore AI documents with a potential contribution to sustainable development, we performed three analyses: The first one compares the number of documents by IEEE taxonomy categories and documents within those categories related to sustainable development. This intends to answer our first research question: *To what extent is sustainable development addressed in AI research produced globally?*

The second analysis responds to question 2, *Which subject areas of AI show an emerging interest in sustainable development?* It identifies those IEEE categories that are having a critical growth rate in three aspects: the number of documents produced, citation impact, and inventive impact. Citation impact is the sum of the number of citations received by the category from other bibliographic outputs, and inventive impact is the sum of citations of the category from patents. The indicators are defined as:

$$1)\ \text{Publications} = \ CAGR_{Publications} = \left(\frac{P2019}{P2000}\right)^{\left(\frac{1}{2019-2000}\right)} - 1 * 100$$

$$2)\ \text{Citation impact} = \ CAGR_{Citations} = \left(\frac{C2017}{C2000}\right)^{\left(\frac{1}{2017-2000}\right)} - 1 * 100$$

$$3)\ \text{Inventive impact} = \ Citations\ Pat\ per\ document = \frac{CP_{Total}}{P_{total}}$$

Where CAGR = Compound Aggregate Growth Rate; Publications = is number of documents; Citations = number of citations from bibliographic outputs; and Citations Pat = citations from patents.

Concerning the citation impact indicator, we take as the most recent year 2017 because citations are delayed accumulating and may show a decrease that is not accurate. For patent citations, we provide the patent citations per document, as this indicator does not behave linearly and thus does not allow us to forecast the linear trend expected by CAGR.

Additionally, we calculated a composite indicator based on the average ranking of the three variables. A comparison between sustainable development-related production and other production is offered.

In addition to the previous analyses, we performed a third analysis to address the research question *To what extent does AI stimulate collaboration between regions of the world?* Given that there is competition among nations for leadership in this field (Mou 2019), it can be expected that cross-regional collaboration is minimum, as well as collaboration between countries classified in different income groups. We also offer a comparison between economic regions in terms of the World Bank's income categories. The indicators for performing these comparisons rely on counting papers once for each author/region. It produces the same number of total papers if there is no collaboration (authors of the papers belong to the same region). Otherwise, the number could be greater than the number of papers, given that collaboration from two authors of different regions could count as 1 for each region. The indicators are as follows:

1) Rate of interregional collaboration = Sum of documents produced by region / total documents
2) Rate of cross-income collaboration = Sum of documents produced by income category/total of documents

Finally, we performed a social network analysis of countries to identify the most central nodes through co-authorships. The literature on research collaboration at the country level is consistent in the finding that a few countries concentrate a huge amount of power in the networks, while many others stay at the periphery of collaboration (Kwiek 2020; King 2011; Wagner 2008). This gives those central countries the power to control knowledge flows and the rules of collaboration (Olechnicka et al. 2019). We explored the collaboration structure of AI papers related to SD by analyzing their betweenness centrality, which is a measure based on the number of shortest paths that cross through the nodes of the network.

Results

Consideration of sustainable development in AI research

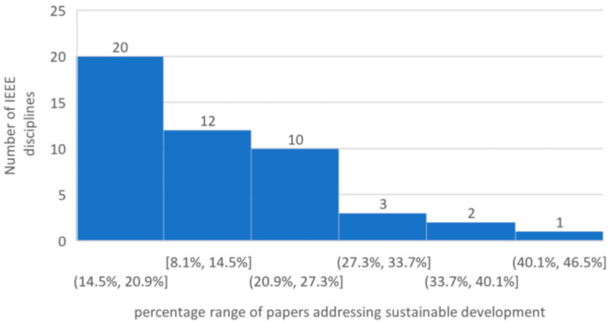


Figure 1. Distribution of the number of disciplines by the percentage of documents addressing sustainable development.

Source: Authors' elaboration based on IEEE Xplore and taxonomy

Figure 1 shows the number of documents by IEEE taxonomy category and the percentage of publications that address sustainable development. In total, 197,384 documents had an IEEE taxonomy classification. Of these, 44,563 papers were classified as sustainable development, which means that approximately 23% of papers in our dataset addressed sustainable development, based on our keyword identification. Figure 1 shows a Pareto distribution of these categories. Most of the disciplines address sustainable development between 14% and 27%. Only five disciplines address sustainable development in more than 27% of their papers, with only one reaching the limit of 40%. This hints that sustainable development is a subject highly concentrated in specific domains of AI.

IEEE subjects with an emerging interest in sustainable development.

The indicators for total AI documents, AI documents related to sustainable development, and AI documents not classified as sustainable development are in Table 3. The calculations include production CAGR, scientific impact CAGR (based on citations from bibliographic outputs), and invention impact (based on patents), as outlined in the methodology. In general, we find that sustainable development documents supersede other AI documents in production growth and scientific impact. This shows that papers in AI are increasingly addressing sustainable development subjects and that the citation rate to articles addressing sustainable development is increasing above-average citation growth. However, sustainable development documents perform below average on inventive impact as measured by patent citations per document.

Table 3. Production, scientific impact, and inventive impact indicators for AI papers for the analyzed database.

	Production (CAGR)	Scientific Impact (CAGR)	Inventive impact (patent cites per doc)
All docs	15%	7%	0.15
SD docs	22%	14%	0.11
Non-SD docs	14%	7%	0.16

Note: All docs refer to the totality of AI documents; SD docs point to AI documents related to sustainable development; Non-SD docs relate to AI documents not classified as related to sustainable development.
Source: own calculations based on IEEE Xplore and taxonomy

Table 4. Top 5 emerging subjects addressing sustainable development in AI papers

IEEE subject	CAGR Prod %	CAGR Cit. %	Pat ind.	Rank Prod.	Rank Cit.	Rank Pat	Avg rank
Broadcast technology	34.3	44.1	0.007	2	5	17	8
Systems engineering and theory	33.9	29.7	0.019	3	17	6	8.67
Ultrasonics, ferroelectrics, and frequency control	33.8	51.5	0.004	4	2	22	9.33
Sensors	32.0	36.8	0.010	9	9	12	10
Education	41.5	64.0	0.000	1	1	33	11.67

Note: CAGR stands for Cumulative Aggregate Growth Rate. Prod stands for documents, Cit. stands for the sum of citations, Pat ind refers to patent citations per paper. Rankings with lower values must be understood as being at the top of the list. Each ranking refers to the ordering of the values of CAGR and Pat ind. The average ranking is the mean of the three rankings in the columns.

The above indicators allowed us to derive a ranking of emerging areas in AI and sustainable development. The ranking was calculated as the average of the rankings in the three above indicators for each IEEE subject area. It identifies the subjects that show a more pronounced growth in the number of documents and citations in bibliographic outputs, and a higher score in patent cites per document. Table 4 shows the top 5 subjects.

Table 5 shows the correlation coefficients between the three rankings above. The correlations show that production growth is moderately and positively associated with scientific impact as measured by citation growth. However, inventive impact as measured by patent citations per document is weakly related to production and scientific impact. This means that even if sustainable development papers increase their production and citation growth, this does not necessarily indicate that sustainable development-related knowledge will make its way into the development of technologies.

Table 5. Correlation coefficients between the ranking of production, the ranking of scientific impact, and the ranking of inventive impact.

	Production Ranking	Scientific impact ranking	Inventive impact ranking
Production Ranking	1	0.6	0.3
Scientific impact ranking		1	0.1
Inventive impact ranking			1

Collaboration in AI papers related to sustainability between regions, income groups, and countries.

Production of AI papers indexed in IEEE Xplore is highly concentrated in high-income countries from the East Asia & Pacific, Europe & Central Asia, and North America regions. Tables 6 and 7 show the distribution of papers by region and income group.

Table 6. Number of AI documents by region

Region	Docs	Docs - SD	Docs - non-SD
East Asia & Pacific	100,692	18,694	81,998
Europe & Central Asia	56,188	11,771	44,417
North America	47,738	9,990	37,748
South Asia	16,469	4,808	11,661
Middle East & North Africa	10,191	2,587	7,604
Latin America & Caribbean	7,339	1,687	5,652
Sub-Saharan Africa	1,481	461	1,020

Note: Docs refers to the total number of AI documents; Docs – SD relates to the number of documents related to sustainable development; Docs – non-SD relates to the number of documents not identified as related to sustainable development.

Table 7. Number of AI documents by income group

Income group	Docs	Docs – SD	Docs - non-SD
High income	119,397	24,263	95,134
Upper middle income	95,163	18,355	76,808
Lower middle income	21,173	6,077	15,096
Low income	203	64	139

Note: Docs refers to the total number of AI documents; Docs – SD refers to the number of documents related to sustainable development; Docs – non-SD refers to the number of documents not identified as related to sustainable development.

In general, inter-regional collaboration in AI papers between countries is low. The ratio between the total number of papers and co-authorships between regions is 1.18. Papers related to sustainable development have a lower ratio of 1.12. Papers not identified as related to sustainable development have a ratio of 1.18. This means that most papers involve only collaboration within their region.

In terms of income group, the data shows similar patterns. The ratio between the total number of papers with a country of affiliation identified and co-authorships between income groups is 1.16. Papers related to sustainable development and papers not related to sustainable development have the same ratio. As in the regional case, this means that most papers involve only collaboration within their income group.

Figure 2 shows the social network graph of country co-authorships for SD-related papers. Node size and color intensity indicate the betweenness centrality of countries, and link thickness represents the number of co-authorships between countries. The top 10 percentile by betweenness centrality is composed, in descending order, by the United States, France, China, United Kingdom, India, Spain, Japan, Austria, Saudi Arabia, Germany, Malaysia, Australia, Portugal, Croatia, and Russia. Of these fifteen countries, eleven are high-income countries, three are upper-middle income, and one (India) is classified as a low-income country. Although the top 10 percentile is dominated by high-income countries, the general correlation between betweenness centrality and income group is only 0.21, which shows that there is a diversity of countries that revolve around the most central ones. The social network analysis partly confirms the concentration of power in a few countries pointed out by the literature on collaboration and shows the emergence of new powers in SD-related AI research.

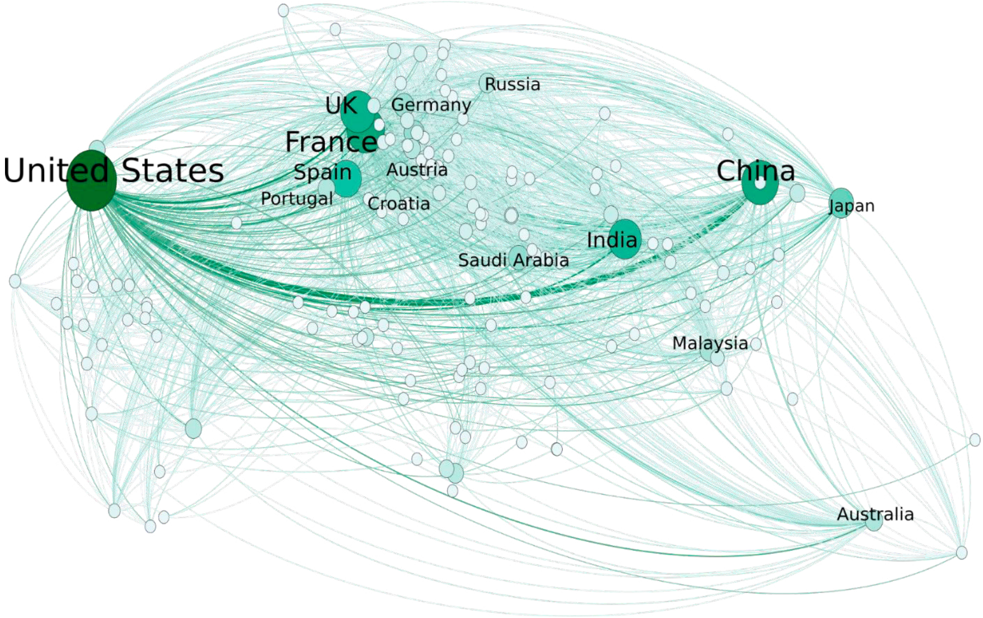


Figure 2. Co-authorship network of countries*

* Nodes positions attempt to preserve the location of countries in the Mercator projection of the world map. Size and color indicate betweenness centrality; only countries in the top 10 percentile by betweenness centrality shown for visualization.

Discussion and conclusion

In this research we addressed the questions: 1) To what extent is sustainable development addressed in AI research produced globally? 2) Which subject areas of AI show an emerging interest in sustainable development? And 3) To what extent does AI stimulate collaboration between regions of the world?

We found that AI papers address sustainable development to a small extent. In most subjects, the percentage of papers addressing sustainable development ranges between 8% and 30%. The most related subjects are P*ower engineering and energy, Geoscience and remote sensing*, *Engineering management*, and *Engineering in medicine and biology.* Even in these subjects we only found between 30% and 40% of papers related to SD. The small number of explicit references to SD in AI papers indicates that discussions on the use of AI for grand challenges such as poverty reduction, inequality, peace, among others, are not central to the engineering AI literature. Motivating engineers to break their silos and explicitly reflect on the connection between their technologies and society will help increase the value of AI research beyond technical progress and efficiency gains.

Global AI research policies could build on our finding that SD is an emerging concern for engineering researchers in certain subjects such as *Broadcast technology*, *Systems engineering and theory*, *Ultrasonics, ferroelectrics, and frequency control*, *Sensors,* and *Education.* Understanding in which ways AI is being linked to SD by engineers in these and other fields could help to design strategies to make AI more relevant for grand societal challenges. Engineering education is one key field in which interdisciplinary research about sustainable development could have a profound impact, given the cross-cutting nature of education to the development of all engineering fields.

In terms of collaboration, our findings show that inter-regional and inter-income group collaboration is not very frequent. The data shows that 1 in 7 papers will produce an inter-regional or inter-income group collaboration. This pattern may be a reproduction in research of the competitive environment for leading technological development, as pointed out by the literature (Mou 2019; Sørensen et al. 2016). We also see an enormous concentration of research capacity in East Asia & Pacific, Europe & Central Asia, and North America. This shows a gap in knowledge production that confirms the technological and knowledge dependence of certain regions and reinforces the global division between centers and peripheries (Kshetri 2020). We see that this competitive and monopolistic environment is likely to produce an even greater gap between technologically advanced countries and the rest, which goes in contradiction to the stated goals of global AI policies and SD.

Based on our experience in policymaking in an upper middle-income country, we know that the technological gap will not be closed if countries at the periphery continue to be seen as mere buyers of technology and their governments continue to accept this role. Changing this mindset requires a committed public investment in R&D to create and improve local capacity to the extent that a country can be a producer and not only a consumer of AI. While being a producer will provide an opportunity to define more relevant research agendas related to the specific needs of countries and regions, being a consumer will continue to increase subordination (Dussel 2018). India and China have shown that a clear commitment to the development of local research capacity in AI is beneficial in the long term. However, individual countries cannot by themselves change the global landscape, because even if they started investing in R&D, local capacity takes time to build. If global AI policies are to support SD, and SD is a global agenda, then there should be a global commitment to build AI capacity in those countries that remain peripheral, instead of concentrating power and capacity on a few ones. This requires a change towards a more generous attitude in international relationships that is not supported by the current climate of competition between countries such as the USA, Russia, and China, and by the current rate of natural resources extraction.

The above findings are not only relevant for research policy, but also SD as a societal goal. Increasing awareness of researchers about SD and their motivation to produce socially relevant research, as well as increasing the capacity to produce such research is determining for solving some of the grand challenges of the world. More aware and motivated researchers will understand the need to collaborate with stakeholders beyond academia to produce technologies that matter for their societies, while developing new technical and social skills that will build the capacity of countries to determine and pursue their own research agendas. This is aligned with the stated aims of multilateral organizations such as the OECD and the

European Union, as well as institutions such as the IEEE, which have expressed a determined intention to produce AI research for the benefit of the economy, the environment, and society. They have produced formal statements to support this orientation of AI to pave the way to that goal. Our results offer an empirical basis for these organizations and for countries to connect AI engineering research with society. Sustainable development provides a framework to achieve this connection, which can be compared to what Jane Goodall calls a connection between our brain -committed to advancing the frontier of science- and our heart, that connects scientific advancement to the well-being of the planet.

Future research

This is an exploratory analysis that opens different avenues of research, for instance in deepening our analysis on collaboration, interdisciplinarity, and emerging subjects. Further research should also investigate the social, economic, and environmental contextual dimensions of the countries that produce AI research. On the technical side, our research can be enhanced by extending the analysis of AI through different classification standards beyond IEEE and including discussions of Artificial Intelligence interpretations as discipline and as a research field.

Acknowledgments

This work was supported by the Ministry of Science, Technology, and Innovation of Colombia (Minciencias) under Grant 848-2019 and the Colombian Society of Engineering Physics (SCIF), Colombia. We thank the IEEE for offering their open API, OSDGs for their web service to classify papers through machine learning, especially to Gusté Statulevičiūtė and Lukas Pukelis from PPMI. We also thank the Free and Open-source software community: MariaDB, R, Gephi, and Notepad++ teams for making available their software. Special thanks to an anonymous reviewer and to Professor Wolfgang Glänzel for their excellent suggestions.

References

Armitage, C. S., Lorenz, M., & Mikki, S. (2020). Mapping scholarly publications related to the Sustainable Development Goals: Do independent bibliometric approaches get the same results?. *Quantitative Science Studies*, *1*(3), 1092-1108.

Breschi, s., Lissoni, F., Montobbio, F. (2005). From publishing to patenting. Do productive scientists turn into academic inventors? *Revue d'économie industrielle*, (110), 75-102

Cassi, L., Lahatte, A., Rafols, I., Sautier, P., & De Turckheim, E. (2017). Improving fitness: Mapping research priorities against societal needs on obesity. *Journal of Informetrics*, *11*(4), 1095-1113.

Chavarro, D., Vélez, M. I., Tovar, G., Montenegro, I., Hernández, A., & Olaya, A. (2017). *Los Objetivos de Desarrollo Sostenible en Colombia y el aporte de la ciencia, la tecnología y la innovación*. Colciencias. https://minciencias.gov.co/sites/default/files/ctei_y_ods_-_documento_de_trabajo.pdf

del Río Castro, G., Fernández, M. C. G., & Colsa, Á. U. (2020). Unleashing the convergence amid digitalization and sustainability towards pursuing the Sustainable Development Goals (SDGs): A holistic review. *Journal of Cleaner Production*, *280*(122204).

Dussel, E. D. (2018). Hacia la liberación científica y tecnológica en América Latina. *Vectores de investigación*, (14), 103-112.

European Commission. (2020). *Emerging trends in STI policy*. European Commission. https://stiplab.github.io/R2r/Emerging%20trends%20in%20STI%20policy.html

European Parliament. (2020). *Ethics of artificial intelligence.* European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

European Patent Office. (2020). *Patents and the fourth industrial revolution*. European Patent Office. http://documents.epo.org/projects/babylon/eponet.nsf/0/06E4D8F7A2D6C2E1C125863900517B88/$File/patents_and_the_fourth_industrial_revolution_study_2020_en.pdf

Elsevier (2020). *Mapping research to advance the SDGs*. Elsevier. https://www.elsevier.com/connect/sdg-report

Goodall, J. (2014). *Interview*. Brainpickings. https://www.brainpickings.org/2014/09/30/jane-goodall-empathy/

Harari, Y. N. (2017). Reboot for the AI revolution. *Nature News*, *550*(7676), 324.

IEEE. (2019). *IEEE position statement on artificial intelligence*. IEEE. https://globalpolicy.ieee.org/wp-content/uploads/2019/06/IEEE18029.pdf

IEEE. (2021a). *IEEE Thesaurus*. IEEE. https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/ieee-thesaurus.pdf

IEEE. (2021b). *Artificial Intelligence Systems (AIS) Related Standards*. IEEE. https://standards-ieee-org.ezproxy.utp.edu.co/initiatives/artificial-intelligence-systems/standards.html

Kagermann, H., Anderl, R., Gausemeier, J., Schuh, G., & Wahlster, W. (Eds.). (2016). *Industrie 4.0 in a Global Context: strategies for cooperating with international partners*. Herbert Utz Verlag.

King, R. (2011). Power and Networks in Worldwide Knowledge Coordination: The Case of Global Science. *Higher Education Policy 24*(3), 359–76.

Klavans, R., Boyack, K. W., & Murdick, D. A. (2020). *A Novel Approach to Predicting Exceptional Growth in Research*. ArXiv. *arXiv:2004.13159*.

Kshetri, N. (2020). Artificial Intelligence in Developing Countries. *IEEE Annals of the History of Computing*, *22*(04), 63-68.

Kwiek, M. (2020). What large-scale publication and citation data tell us about international research collaboration in Europe: changing national patterns in global contexts. *Studies in Higher Education*, ahead of print, 1-21.

Mou, X. (2019). *Artificial Intelligence: Investment Trends and Selected Industry Uses*. World Bank. http://documents1.worldbank.org/curated/en/617511573040599056/pdf/Artificial-Intelligence-Investment-Trends-and-Selected-Industry-Uses.pdf

Navarro, J. (2018). *The digital transformation imperative. An IDB science and business innovation agenda for the new industrial revolution.* IDB. https://publications.iadb.org/en/digital-transformation-imperative-idb-science-and-business-innovation-agenda-new-industrial

Nilson, N. (2010). The quest for artificial intelligence. *A history of ideas and achievements.* Cambridge University Press.

OECD. (2020). *Principles of AI*. OECD. http://www.oecd.org/going-digital/ai/principles/

Olechnicka, A., A. Ploszaj, & D. Celinska-Janowicz (2019). *The Geography of Scientific Collaboration*. Routledge

OSDG. (2021). *Home page*. OSDG. https://osdg.ai/

Osoba, O. A., & Welser, I. V. W. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation. https://www.rand.org/pubs/research_reports/RR1744.html

Planes-Satorra, S. & Paunov, C. (2019). *The digital innovation policy landscape*. OECD. https://doi.org/10.1787/6171f649-en

Russell, S. (2010). *Artificial intelligence: a modern approach*. Prentice Hall, 2010.

Schwab, K. 2017. *The fourth industrial revolution.* World Economic Forum.

Sørensen, J., Tadayoni, R., & Henten, A. (2016). 5G-Boundary Object or Battlefield?. *Communications & Strategies*, (102), 63-87.

UNIDO. (2018). *Industry 4.0 – The opportunities behind the challenges.* Unido. https://www.unido.org/sites/default/files/files/2018-11/UNIDO_GC17_Industry40.pdf

Wagner, C. (2008). *The new invisible college. Science for development*. Brookings Institution Press.

Zellmer-Bruhn, M. E., Forbes, D. P., Sapienza, H. J., & Borchert, P. S. (2021). Lab, Gig or Enterprise? How scientist-inventors form nascent startup teams. *Journal of Business Venturing*, *36*(1), 106074.

A Fine-Grained Annotation Scheme for Research Contribution in Academic Literature

Haihua Chen[1] and Bhavya Nandana Kanuboddu[2]

[1] *haihua.chen@unt.edu*
Department of Information Science, University of North Texas, Denton, Texas (USA)

[2] *BhavyaNandanaKanuboddu@my.unt.edu*
Department of Information Science, University of North Texas, Denton, Texas (USA)

Abstract

Research contributions are the most valuable information which can help researchers understand the main content of academic literature. However, existing research mainly focuses on citation analysis, rhetorical structure analysis, essential sentence extraction, and little attention is paid to the research contributions stated in the full text. This paper first introduces a fine-grained annotation scheme with nine categories for research contributions in academic literature. To evaluate the reliability of our annotation scheme, we conduct the human annotation on 3,104 sentences collected from 3,293 articles in ACL Anthology and achieve an inter-annotator agreement of 0.89 regarding the Kappa measurement. We analyze the research contributions in different types and years, showing that new methods, new models, and new theories are the most welcomed research contributions of the ACL conference. In the meanwhile, the percentage of different contributions is stable over the years. The research contribution dataset developed in this paper provides the basis for the automatic research contributions extraction and knowledge fragment recommendation. The fine-grained annotation scheme can be applied for large-scale analysis for research contributions in academic literature.

Introduction

Academic literature records the research process with a standardized structure and provides clues to track the progress in a scientific field (Fisas et al., 2016). Generally, the main components of academic literature include an abstract, introduction, related work, method, experiment and result, conclusion (Lu et al., 2018). In recent years, academic text mining using content from different components has received increasing attention from researchers. However, most of the existing research focuses on keyphrase extraction (Park & Caragea, 2020), citation content analysis (Fisas et al., 2016), rhetorical structure analysis (Sateli & Witte, 2015), essential sentence extraction (Mehta et al., 2018), and little attention is paid to the research contributions stated in the full text. Research contributions, known as the most valuable information, are required to be highlighted in the introduction when a paper is published. A research contribution relates to the research problem addressed by the contribution, the research method, and (at least one) research result (Oelen et al., 2019). For example, 'we build a transfer learning framework employing a diverse range of intermediate tasks covering sequence tagging with semantic and syntactic aspects, and natural language inference' and 'we achieve competitive performance over both strong baselines and previous works' are two contributions stated in Park & Caragea (2020). Research contributions can help researchers understand the core content of a paper and the innovation growth of science.

We can easily identify sentences about research contributions from the introduction section by following the statements like 'Our contributions are summarized as follows', 'The major contributions of this paper are', or similar statements. There are different types of contributions, for example, creating datasets, building new models, performing evaluations, etc. If these contributions can be classified appropriately, it would be helpful for knowledge recommendation, structured abstract generation, scientific evolution analysis. However, an annotation scheme for research contributions is one of the essential requirements for research contribution classification.

We studied the existing annotation schemes regarding academic literature. We noted that most of them mainly focus on context types (Angrosh et al., 2012), citation functions (Teufel et al., 2006), term functions (Li et al., 2017), and future work types (Hao et al., 2020). There is no annotation scheme for research contributions. To bridge this gap, we propose a fine-grained annotation scheme with nine categories for research contributions in academic literature. The human annotation experiment conducted on 3,104 sentences collected from 3,293 articles in ACL Anthology demonstrates the reliability of our scheme. The final annotated research contribution corpus is available for download at: https://zenodo.org/record/4483757#.YBbnw-j0mbg.

The contributions of our paper are as follows:

- We propose a fine-grained annotation scheme for research contributions in academic literature.
- We conduct the human annotation to evaluate the reliability of our scheme.
- The corpus developed in our paper provides basic training data for automatic research contributions classification.

Related Works

This paper proposes an annotation scheme for research contributions in academic literature. Therefore, the related works include research contributions analysis and annotation schemes constructed for academic literature.

Research Contributions Analysis

Research contributions analysis is a new topic, which was paid attention to recently. The Open Research Knowledge Graph (ORKG) constructed a contribution knowledge graph where each paper was summarized with its fundamental contribution properties and values (Auer et al., 2018). In ORKG, the contributions were interconnected via the graph, even across papers. It helps users to compare research contributions between different papers while writing an academic literature review (Oelen et al., 2019). Similarly, Vogt et al. (2020) represented research contributions in scholarly knowledge graphs using knowledge graph cells. Compared to ORKG (Auer et al., 2018), the Research Contribution Model (RCM) can generate a KG whose content is better and maintainable and easier understandable (Vogt et al., 2020). The ontology built by Vogt et al. (2020) provided a reference for defining the annotation categories in our study. However, it identified contributions from abstracts rather than full texts. D'Souza &Auer (2020) developed an annotation scheme to identify research contributions from NLP literature with the structure of <subject, predicate, object>. More recently, Le et al. (2019) applied research contributions identified from citing papers for evaluating the academic value of cited papers. Zhou et al. (2020) extracted academic innovation and contributions from full texts in the field of chrysanthemums using BERT. Although existing research has made certain progress in research contributions identification and representation, an annotation scheme is still needed for fine-grained analysis of research contributions.

Annotation Schemes for Academic Literature

Several annotation schemes have been proposed for academic literature. The annotation schemes are either designed for full texts or citation sentences only. Regarding full texts level, Hao et al. (2020) proposed an annotation scheme that included six main categories and 17 sub-categories for future work sentences. D'Souza &Auer (2020) described ten core information units for organizing academic contributions data in a knowledge graph, including ResearchProblem, Approach, Objective, ExperimentalSetup, Results, Tasks, Experiments, AblationAnalysis, Baselines, and Code. As for the citation sentence level, Teufel et al. (2006) proposed an annotation scheme with four categories and 12 fine-grained categories for citation function. The annotation experiment was performed on 320 conference articles and the agreement calculated with Kappa was used to measure the reliability. Differently, Angrosh et al. (2012) presented a citation-centric annotation scheme for academic literature. It included six categories for citation sentences and five categories for non-citation sentences. A pilot study was carried out using 11 annotators and nine articles. Agreement calculated with Krippendorff's Alpha was used to measure the reliability. The above research provides us insight into how to construct a fine-grained annotation scheme for research contribution sentences and evaluate the reliability of our scheme.

Annotation Scheme for Research Contribution

With the increasing of submissions, it is more and more challenging for reviewers to review a manuscript. On the other hand, users might spend much time reading but hardly figuring out valuable information from the paper. Therefore, highlighting the main contributions when submitting a manuscript becoming a requirement for many conferences and journals. A research contribution can be reflected from different aspects, such as creating a new dataset, proposing a new theory or framework, designing a new model, etc. Studies have discussed the fundamental concepts of academic literature, which are essential for identifying the type of research contributions. Q. Zadeh & Handschuh (2014) defined seven core conceptual classes for Computational Linguistics publications: Technology and Method; Tool and Library; Language Resource; Language Resource Product; Models; Measures and Measurements; and Other. D'Souza & Auer (2020) defined ten core concepts for research contributions for NLP publications: Research Problem, Approach, Objective, Experimental Setup, Results, Tasks, Experiments, Ablation Analysis, Baselines, and Code. The annotation scheme proposed for future work sentences in Hao et al. (2020) included six categories and 17 sub-categories: Method (Model, Algorithm, Feature, Other), Resources (Optimize resources, Expand resources, Change resource type, Other), Evaluation (Evaluate current work, improve evaluation means, expand evaluation areas, Other), Application (Applied for other tasks, Expand application areas), Problem, Other.

Our scheme (shown in Table 1) is a combination of the scheme in Hao et al. (2020), D'Souza & Auer (2020), and Q. Zadeh & Handschuh (2014), which we finalize after an analysis of a corpus of academic literature in computational linguistics. We try to redefine the categories to be adapted for research contributions and reasonably reliably annotatable. Our categories are as follows: one category is Dataset Creation. In the computational linguistics field, creating a new dataset is a common research contribution (Fisas et al., 2016; Hao et al.,2020) because other researchers for similar studies can reuse the new dataset. The second category is Theory Proposal. As we know, science is knowledge represented as a collection of "theories" derived using the scientific method (Somekh & Lewin, 2005). Theories provide explanations of natural or social behavior, event, or phenomenon (Somekh & Lewin, 2005). Proposing a theory is considered one of the essential contributions of academic research. The next four categories are Model Construction, Model Optimization, New Algorithm or Method or Technology, and Algorithm or Method or Technology Optimization. Different from the annotation scheme in Hao et al. (2020), we separate the Method and Model since they are different concepts, especially in computational linguistics, according to Q. Zadeh & Handschuh (2014). Methods, algorithms, and technologies are designed, developed, and employed to accomplish a specific task to fulfill a practical purpose. Simultaneously, models represent what is learned by an algorithm based on the data (Q. Zadeh & Handschuh, 2014). Both models and methods can be constructed from scratch or optimized based on the existing models and methods. Therefore, we distinguish research contributions from model construction, model optimization, method construction, and method optimization. As for the last three categories, including Performance Evaluation, Resources, and Application, we reuse and adapt from Hao et al. (2020).

Table 1. Our annotation scheme for research contributions.

Category	*Description*	*Example*
Dataset Creation	Build a new dataset	We propose to build multi-task datasets for the News and Tweets do-mains, by unifying the a forementioned task-independent datasets.
Theory Proposal	Propose a new theory to solve existing problems or for improvisations	We suggest viewing learning event embedding as a multi-relational problem, which allows us to capture different aspects of event pairs.
Model construction	Construct a model	We propose to generate comments with a graph-to-sequence model that models the input news as a topic interaction graph.
Model Optimization	Propose strategies to increase the efficiency of existing model	We propose an approach to improving the robustness of NMT models, which consists of two parts: (1) attack the translation model with xxx; (2) defend the translation model with xxx.
New Algorithm or Method or Technology	Develop a new method or algorithm or technology	We present a new method for sentiment lexicon induction that is designed to be applicable to the entire range of typological diversity of the world's languages.
Algorithm or Method or Technology Optimization	Optimize the existing method or algorithm or technology	Our submission combined techniques including utilization of a synthetic corpus, domain adaptation, and a placeholder mechanism, which significantly improved over the previous baseline.
Performance Evaluation	Evaluate the performance of the new or existing implementation	We evaluate a broad variety of neural models on the new dataset, establishing strong baselines that surpass previous feature-based models in three tasks.
Resources	Expand/change the types of resources	We provide a resource to be distributed for research purposes in the BioNLP community.
Applications	Apply the proposed model or theory to other tasks	Our model can also be applied to the cross-domain named entity recognition task, and it achieves better adaptation performance than other existing baselines.

Data

The data we use comes from ACL Anthology (https://www.aclweb.org/anthology/venues/acl/). We collect 3,293 articles from the proceedings of the Annual Meeting of the Association for Computational Linguistics, out of which 674 are from 2018, 1393 are from 2019, and 1226 are from 2020. The articles are in pdf format. We manually identify the sentences which indicate the research contributions. Specifically, we conduct a pre-investigation of the original corpus to summarize the patterns of research contribution sentences, then formulate the labeling specifications. For explicit research contribution sentences, we can easily identify by following the contribution block indicators, for example, 'Our contributions are summarized as follows', 'The major contributions of this paper are', or similar statements. As for implicit contribution sentences, we find that their locations in the paper are usually the last paragraph or the second last paragraph in the introduction section. We then figure out several verbs or verb phrases that indicate the research contributions, such as the present, introduce, compare, design, apply, develop, etc. By following the above strategy, we collect 3,104 research contribution sentences in total.

Annotation Experiment and Results

To evaluate the reliability of the research contribution annotation scheme discussed before and create a dataset for automatic research contribution classification, we conduct the annotation experiment with six annotators who are designers of the scheme and very familiar with the annotation guideline. They also have a background in NLP and machine learning, which can ensure annotation quality.

Annotation Results

The six students are divided into pairs for the annotation; in other words, each sentence will be annotated by two annotators. During the annotation, the annotators independently annotated the same number of sentences with the proposed scheme. We measure the agreement in Kappa, which ranges between -1 and 1. Generally, Kappa's of 0.8 is considered stable, and Kappa's of 0.69 as marginally stable (Carletta, 1996). We reached an inter-annotator agreement of K = 0.89. We consider the agreement quite good, considering the number of categories. For sentences that are annotated with different categories by the two annotators, a third annotator will perform the annotation, and a majority vote will be used to decide the final category.

Statistical Analysis of the Annotated Dataset

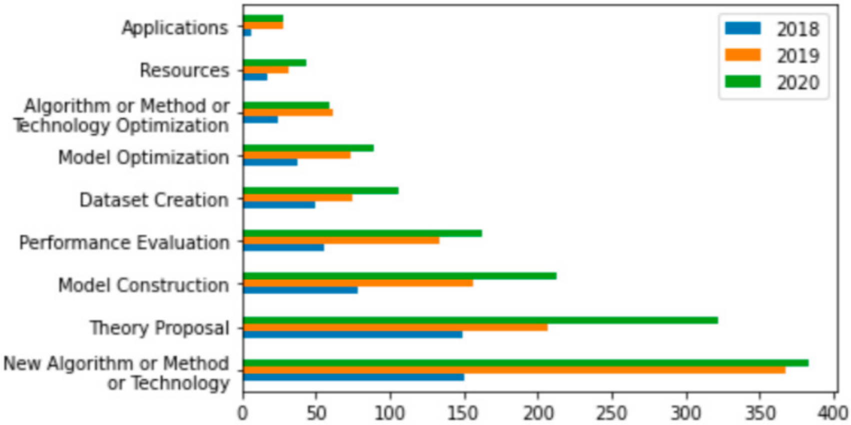
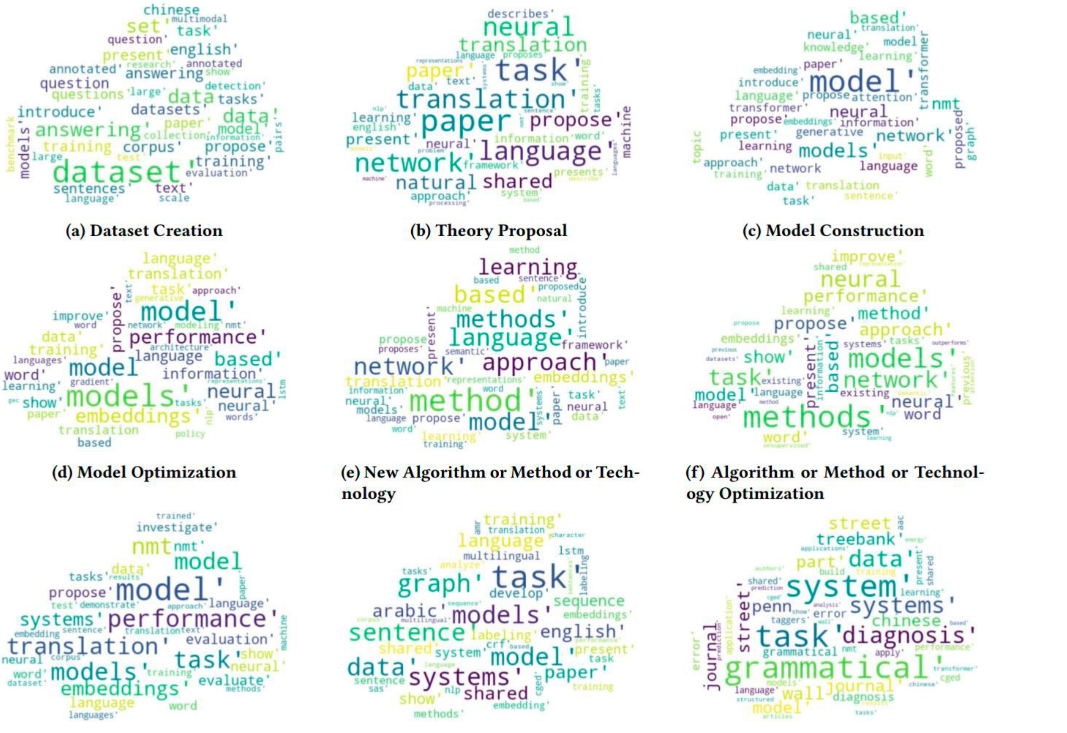


Figure 1: The distribution of research contributions in each category over years

The relative frequency of each category for each year observed in the annotation results is shown in Figure 1. The distribution is very skewed, with 27.1% of the research contributions of the category New Algorithm or Method or Technology. Interestingly, the relatively high frequency of the New Algorithm or Method or Technology, Theory Proposal, Model Construction, Performance Evaluation categories with 76.6% in total. As we have known, the four categories are more substantial contributions than the rest five categories in computational linguistics—this concordance with the current research directions and concentrations on high-quality computer science conferences (https://mohammadkhalifa.github.io/2020/07/30/ACL2020-highlights.html). With the dramatic increase of submissions, top conferences require submissions to demonstrate stronger contributions for being accepted. As for the distribution of research contributions over the years, we find similar patterns.

Keyword Analysis on Each Category

We analyze the content of all research contribution sentences using YAKE! (Campos et al., 2020), to extract the keywords. YAKE! is the state-of-the-art tool for keyword extraction. It defines a set of features (casing, position, frequency, relatedness to context, and dispersion of a specific term), capturing keyword characteristics that are heuristically combined to assign a single score to every keyword (Campos et al., 2020). We extract the top 40 keywords (as listed in Figure 2) for the contribution sentences in each category and the keywords are limited to two grams.



(a) Dataset Creation

(b) Theory Proposal

(c) Model Construction

(d) Model Optimization

(e) New Algorithm or Method or Technology

(f) Algorithm or Method or Technology Optimization

(g) Performance Evaluation

(h) Resources

(i) Applications

Figure 2: The word cloud of the research contribution sentences in each category (top 40 keywords)

As shown in Figure 2, the keyword distribution reflects the most important contributions in each category and the verbs used to present the research contributions in each category. For example, in the category Dataset Creation, it is obvious that the keywords "data", "dataset", "corpus", "xxx dataset" are at the top of the list. Verbs such as "present" and "introduce" are frequently used to introduce a new dataset. Similarly, in the Model Construction and Model Optimization categories, keywords such as "model", "neural network model", "language model" are prevalent. However, the keyword "improve" appears in the Model Optimization category but does not appear in the Model Construction because "improve" indicates that academic literature contributes by optimizing previous models rather than introduce a new model. A more detailed study of the keywords may be helpful for automatic research contribution classification in the future.

Summary and Future Work

In this paper, we have described a fine-grained annotation scheme for research contribution in academic literature. Our annotation scheme concentrates on identifying different types of research contributions, which can help researchers understand the core content of a paper and the innovation growth of science. We report positive results in terms of the inter-annotator agreement in the annotation experiment. In addition, we conduct statistical analysis and keywords analysis on the constructed research contribution corpus.

We are currently investigating how well our scheme will work on academic literature from a different discipline, namely chemistry. Based on the investigation results, we would be able to optimize the categories so that the annotation scheme can be more widely used. Work on applying semi-automatic approaches to argument the amount of labeled data for automatic research contribution classification is another concentration. Besides, it would be interesting to compare the distribution of research contributions in each category in different fields.

Acknowledgement

The authors would like to thank Aakansha Tallapally, Durga Bhavana Yerrabelli, Nikitha Malladi, and Riyaz Ahmad Shaik for participating the annotation experiment.

References

Angrosh, M., Cranefield, S., & Stanger, N. (2012). A citation centric annotation scheme for scientific articles. *In Proceedings of the Australasian Language Technology Association Workshop* (ALTA'12) (pp. 5-14).

Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., & Vidal, M. E. (2018). Towards a knowledge graph for science. *In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics* (pp. 1-6).

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509, 257-289.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22, 249-254.

D'Souza, J., & Auer, S. (2020). Nlpcontributions: An annotation scheme for machine reading of scholarly contributions in natural language processing literature. *In Proceedings of the 1st Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE'20) co-located with the ACM/IEEE Joint Conference on Digital Libraries* (JCDL'20).

Fisas, B., Ronzano, F., & Saggion, H. (2016). A multi-layered annotated corpus of scientific papers. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16) (pp. 3081-3088).

Hao, W., Li, Z., Qian, Y., Wang, Y., & Zhang, C. (2020). The acl fws-rc: A dataset for recognition and classification of sentence about future works. *In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (JCDL'20) (pp. 261-269).

Le, X., Chu, J., Deng, S., Jiao, Q., Pei, J., Zhu, L., & Yao, J. (2019). Citeopinion: Evidence-based evaluation tool for academic contributions of research papers based on citing sentences. *Journal of Data and Information Science*, 4, 26-41.

Li, X., Cheng, Q., & Lu, W. (2017). Cs-las: A scientific literature retrieval and analysis system based on term function recognition (tfr). *In ISSI* (pp. 680 1346-1356).

Lu, W., Huang, Y., Bu, Y., & Cheng, Q. (2018). Functional structure identification of scientific documents in computer science. *Scientometrics*, 115, 463-486.

Mehta, P., Arora, G., & Majumder, P. (2018). Attention based sentence extraction from scientific articles using pseudo-labeled data. *arXiv preprint arXiv:1802.04675*.

Oelen, A., Jaradeh, M. Y., Farfar, K. E., Stocker, M., & Auer, S. (2019). Comparing research contributions in a scholarly knowledge graph. *In CEUR Workshop Proceedings 2526 (2019)* (Vol. 2526, pp. 21-26). Aachen: RWTH Aachen.

Park, S., & Caragea, C. (2020). Scientific keyphrase identification and classification by pre-trained language models intermediate task transfer learning. *In Proceedings of the 28th International Conference on Computational Linguistics* (ICCL'20) (pp. 5409-5419).

Q. Zadeh, B., & Handschuh, S. (2014). The ACL RD-TEC: A dataset for benchmarking terminology extraction and classification in computational linguistics. *In Proceedings of the 4th International Workshop on Computational Terminology (Computerm)* (pp. 52-63).

Sateli, B., & Witte, R. (2015). What's in this paper? combining rhetorical entities with linked open data for semantic literature querying. *In Proceedings of the 24th International Conference on World Wide Web* (WWW'15) (pp. 1023-1028).

Somekh, B., & Lewin, C. (2005). Research methods in the social sciences. London, UK: Sage.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). An annotation scheme for citation function. *In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue* (pp. 80-87).

Vogt, L., D'Souza, J., Stocker, M., & Auer, S. (2020). Toward representing research contributions in scholarly knowledge graphs using knowledge graph cells. *In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (JCDL'20) (pp. 107-116).

Zhou, H., Zheng, D., & Li, T. (2020). Research on the identification of academic innovation contributions of full academic texts. *Journal of the China Society for Scientific and Technical Information*, 39, 845-851.

Measuring radical and incremental innovation of digital products Updates on text analysis of mobile apps

Jin Chen[1], Xiaohu Li[2] and Lele Kang[*,3]

[1]*Chenjin_98@yeah.net*

[2]*wed26926@163.com*

[3]*lelekang@nju.edu.cn*
School of Information and Management, Nanjing University, No.163 Xianlin Rd., Qixia District, Nanjing, Jiangsu (P.R.China)

Abstract

Prior research agrees that the success of digital products depends on their radical and incremental innovation. Despite the existence of various innovation measures in the literature, it is unclear how to measure the innovation output of digital products. This paper applies content analysis and text mining methods to analyze 39,153 update documents of mobile apps and estimate their radical and incremental innovation. Bidirectional Encoder Representations from Transformers (BERT) was adopted to classify the innovation output into nine categories (e.g. functional, business, and visual updates) based on comparing the classification performance of various models. According to the radical and incremental innovation level, the apps updates can be clustered into four levels, ranging from low to high. Shopping, travel, and finance apps rank highly on the radical innovation level, and the three categories share a similar innovation pattern. In contrast, game apps implement the strategy of radical innovation in their updates. This paper contributes to extending text analysis to measure the innovation output of mobile apps.

Introduction

In recent years, digital products have played a dominant role in economic development. Millions of entrepreneurs have engaged in developing mobile apps and providing digital services via their products. The rapid development of mobile commerce has brought the mobile apps market into a prosperous era. According to Statista's report, Google Play has 2.87 million apps and the Apple App Store released 1.96 million apps in the third quarter of 2020. On the one hand, the mobile phone is continuing to grow, and coverage of all netizens has basically been achieved. The huge scale of mobile phone netizens provides a solid foundation for various apps to open up the market. On the other hand, smart technology development and the widening of 5G apps have enabled a wider range of scenarios for mobile apps and have spawned more subdivided vertical categories of mobile application products.

Although mobile apps have broad prospects, their market is no longer “a blue ocean.” The Matthew effect can be observed in the fierce competition of mobile apps. In other words, those apps which have first mover advantage accumulate more advantages over time, while those which begin with a disadvantage have a lower and lower probability of succeeding (George & Ferraro, 2016). At present, mobile apps with a high degree of interest are mostly divided among Internet giants, especially Facebook, Google, and Apple. Thus, the essential question for developers is how to gain a competitive advantage in the fiercely competitive market.

One critical solution is innovation, which enables mobile apps to acquire new users or maintain survival. On the one hand, innovation can open new markets for apps and give the apps irreplaceable competitiveness (Zhou, Song, & Wang, 2018). On the other hand, innovation requires the app to quickly respond to the refinement and changes in user needs (Comino, Manenti, & Mariuzzo, 2019). Generally, innovation includes implementing a new or improved product or process that differs significantly from the unit's previous products or processes (OECD & Eurostat, 2018) and commercialization of innovation (Boer & During, 2001; Dewangan & Godse, 2014). Edison et al. (2013) proposed that innovation is divided into four categories: product innovation, process innovation, market innovation and organization innovation. Specifically, product innovation involves creation and introduction of new (technological new or significantly improved) products (Geiger & Cashen, 2002; OECD & Eurostat, 2018), which mainly refers to creating processes from scratch.

However, there is a huge difference between digital product, e.g., mobile apps, and physical products. Apps continuously evolve through updates during their life cycle (Zhou, Song, & Wang, 2018). Most apps gradually gain user recognition and competitive advantages through continuous innovation iterations. Some popular apps see updates as frequently as weekly; others release once or twice per month. Thus, the innovation of an app is not only a process from 0 to 1, but more importantly, a process from 1 to infinity (Nelson, 2009). Knowing how innovative the app is, is critical in achieving a competitive advantage in the market.

This study aims to propose a method to measure the innovation output of mobile apps in third-party mobile apps stores. Specifically, update documents of mobile apps serve as a unique data source to measure the innovation. The update document is the demonstration that depicts the revised activities that the developers have created in the new version of the apps. For example, the developers are required to list "What's New" for app updates in the Apple App Store and Google Play. Analyzing these update documents helps to identify the innovation output of the apps.

To accomplish this, we first divided app innovations into radical and incremental innovations based on innovation theory and developed a classification system. Next, we developed content analysis and machine learning methods in sequence to identify each update point from the update document and accordingly obtained the scores of each app in the dimensions of radical innovation and incremental updates. Moreover, clustering and other methods were utilized to analyze the measurement results in-depth.

Theoretical Underpinnings

Innovation of apps

An update (or upgrade) is the behavior whereby software replaces an installed version of a product with a newer version of the same product (Min Khoo & Robey, 2007). A newly updated version may include many new features or a revised appearance. For instance, the 10.2.3 version of Alipay includes two update points, namely optimizing the scan function and adding a "traffic" entry on the homepage. In this study, every update point is regarded as one output of innovation of the app. Previous explorations into the update of apps have been mainly based on three dimensions: degree of the update, update frequency, and update type.

Considering the degree of innovation, Christensen (1989) took the lead in defining disruptive technological innovation and incremental technological innovation. Incremental innovation, also named continuous innovation, means that this type of technology enables products or services to improve existing performance in accordance with those aspects that most users in major markets have always valued (Christensen, 2013). Correspondingly, disruptive innovation or radical innovation refers to radically new products that involve dramatic leaps in terms of customer familiarity and use (Christensen, 2013; Veryzer Jr, 1998). There is a consensus among scholars that both types of innovation have a significant impact on the market performance of new products. Incremental innovation can quickly respond to the refinement and changes of user needs, and maintain its unique differences among competitors (De Brentani, 2001), whereas, radical innovation is likely to bring major competitiveness and open up emerging markets (Hang, Neo, & Chai, 2006).

Apps are iteratively updated products, and the concept of radical innovation is not fully applicable to them. Nevertheless, adding a new business function to the front end of the app through the homogenization and re-programmability of digital technologies can introduce a new category of product or service, which is, to some extent, equal to the introduction of new stand-alone products due to changes in the core components. Based on this view, Tian et al. (2020) defined updates that add new features or functionalities to an existing app as app innovation and updates that improve existing supporting functions as incremental updates.

Many previous studies have proven that frequent updates make for better market applicability. Marvin et al. (2016) conducted a study on the update frequency and unveiled that the identified positive effect on frequent updates can be elicited only with functional feature updates and not with technical non-feature updates. In other words, the type of update has a moderating effect on the positive impact of update frequency (Fleischmann et al., 2016).

The classification of app update types is quite diversified, and there is no unified classification system yet. By manually analyzing the texts in the "What's New" section, Stuart Mellroy and his partners labelled 852 different updates in five categories: new content, new feature, improvement, bugfix, and permission changes (McIlroy, Ali, & Hassan, 2016). Furthermore, changes in the mobile platform will lead to a new type of update to the apps (Soh & Grover, 2020), which provides better compatibility with the platform. Actually, app updates are not limited to product innovation. The innovation of business strategies is also a critical means for many apps to optimize marketing performance (especially for e-commerce apps); hence commercial updates have been involved in the classification system by some researchers as well (Tian et al., 2020).

Metrics for app innovation

Metrics for innovation have become a hotspot in the field of innovation research (Dziallas & Blind, 2019). Innovation penetrates into multiple processes such as corporate management, manufacturing, and marketing. Consequently, the innovation indicators are varied. These can be divided into five main categories: determinants of innovation, inputs of innovation, outputs of innovation, the performance of innovation, and activities of innovation (Edison et al., 2013). Edison, Ali, and Torkar (2013) utilized a complete literature review integrated with questionnaires and interviews with practitioners in the software industry to systematically summarize the existing measurement systems into the following modes: R&D-based measures, revenue-based measures, IPR-based measures, count-based measures, as well as process, investment-based and survey measures. R&D-based measures are used for innovation inputs, while revenue-based measures are used for innovation performance. And it is not hard to understand the process, investment-based and survey measures work for estimating organizations' innovative capabilities. IPR-based measures are the major way to measure innovation output. Patent counts, citation-based volumes and patent density are the most widely used metrics. Nonetheless, "patent only covers a certain type of innovation" (Evangelista, Sandven, Sirilli, & Smith, 1998). Considering the high frequency and multiple types of iterative innovation of apps, IPR-based measures tend to ignore a great deal of non-patented inventions and innovations (Kleinknecht, Van Montfort, & Brouwer, 2002). Therefore, IPR-based measures are obviously not the best choices for innovation of apps.

In summary, update count seems to be the most effective metric in measuring apps innovation. However, count-based measures also have significant flaws in that they treat the quality and degree of all innovation outputs the same. In order to overcome this defect, this study applies the theory of radical innovation and incremental innovation to the measurement system (Qiao et al., 2018).

Based on the above theory, we established a new update classification system to assist in distinguishing the degree of innovation by referring to previous studies and analyzing a large number of update documents. More specifically, we defined nine types of updates as follows: functional updates, business updates, visual updates, visual improvements, existing functional improvements, systematic improvements, bug fixes, adaptability adjustment for platform, together with new content and activities. Functional, business, and visual updates are always accompanied by guidance steps within the app for enabling users to adapt. From the perspective of user familiarity, they have made major changes. Therefore, we classify them as radical innovations, while the remaining six types are considered as incremental innovation. In addition, it should be noted that in the ninth category "new content and activities," both the addition of regular activities and the addition of homogeneous content are predictable to users. That is why we categorized it as incremental innovation.

Table 1. Update classification system and related illustrations

Categories	*Subcategories*	*Code*	*Definitions*
Radical Innovation	1. Functional updates	FU	Addition of new features, new systems, new gameplays.
	2. Business updates	BU	Addition of new sales strategies and business expansion, for example, releasing a specific version opens the shopping festival when the platform provides a certain amount of shopping subsidies.
	3. Visual updates	VU	Changes of the overall design style of the app and revisions of the main interface.
Incremental Innovation	4. Visual improvements	VI	Minor changes to the interface and visual optimization, for example, the position of the icons, the addition of new entrances, and optimized special effects.
	5. Existing functions improvements	FI	Improvement of specific functions and gameplays, for example, adjusting the parameter of existing props and skills.
	6. Systematical improvements	SI	Improvement for the stability, security, fluency, size of installation package, or other systematic performance of the app.
	7. Bug fixes	BF	Changes for handling a programming bug/glitch.
	8. Adaptability adjustment for iOS system	IOSA	Improvement of compatibility with newly upgraded iOS systems.
	9. New content and activities	NCA	Addition of regular non-commercial activities and homogenized content, for example, the addition of new characters, new tasks, or new expression stickers.

Data

Information about 4,560 apps from Apple IOS China was gathered and organized. Some apps were dropped from the analysis as they received very few reviews and gained limited attention from users. Finally, we screened 39,153 pieces of update documents from 2,645 apps from

Apple IOS China. Each had effective update documents and at least ten reviews posted by users from January 1, 2019 to December 31, 2019.

As previously mentioned, before formal analysis, update documents should be divided into update points. For the sake of readability and structure, developers usually describe an update point in one line in the update document. Consequently, we separated the update texts by line breaks and treated each line of a text as a potential update point. Furthermore, some apps repeated previous updates in the latest version of the update document. We removed all of the duplicate content by identifying the keywords "recent updates." Finally, a total of 144,783 pieces of text were obtained.

Table 2. Comparison of update texts before and after preprocessing

A translated update document of WeChat (version code: 7.0.4)		*Extracted update points*
Update:		
When sending "live video," you can search for a song as background music.	1	When sending "live video," you can search for a song as background music.
You can post a private message in the "live video" of your friends.	2	You can post a private message in the "live video" of your friends.
Recent updates: Fixes and improvements to some known issues.		

Methodology

The analysis includes three steps: content analysis and coding, text mining and automatic classification, and update-based measuring. Content analysis is a semi-quantitative research method, which is widely applied to the study of update documents of apps (Vaismoradi & Snelgrove, 2019). However, since its work is mainly done manually, this method is only suitable for a small number of samples (Tian et al., 2020). Considering there are over one hundred thousand pieces of texts to be dealt with, utilizing machine learning methods for text mining has become an inevitable option. Therefore, we first built the training data set with content analysis and utilized it as the input data to discipline the classifier. After that, we counted the number of updates of each category of each app and scored them in the dimensions of incremental innovation and radical innovation.

Step 1: Labelling update points by content analysis

The coding scheme was established according to Table 1. It is worth noting that some invalid texts are inevitably recognized as update points during data preprocessing, such as welcome messages, thank yous, event previews, update previews, activity descriptions, function descriptions, developer’s contact information and information for which it is impossible to automatically determine whether it is an update point. These have been placed into another category, “NA,” which means not an update point in this procedure. To summarize, the complete coding scheme includes three primary codes: radical innovation, incremental innovation and not an update, as well as ten secondary codes.

Approximately 10% of the samples, namely 14,450 pieces of text, were randomly extracted for manual coding. One team member created the coding scheme, and the other two were assigned to code the text separately. After three rounds of coding, the Scottrade index of the two coders reached 0.764. Finally, the remaining borderline and difficult cases were resolved by the first author.

Step 2: Automatically labelling update points by text mining

In order to determine the best classification method, a variety of machine learning models were applied to classify the sample set, including FastText, TextCNN, TextRNN, TextRNN-Attention, TextRCNN, DPCNN, Transformer, and BERT. There are some differences in the implementation of these algorithms. To be more exact, the first seven algorithms all require word2vec to convert text data into word vectors in advance, whereas BERT's unit of analysis is not a word but short text. This is the reason we used the pre-training model of Amazon AWS when applying BERT algorithms. In the training process, we adopted a five-fold, cross-validation method, taking 20% of the data as the test set and 80% of the data as the training set, repeated five times. Ultimately, the BERT model performed best, with an accuracy rate of 81.71%, while the accuracy rates of other models were all lower than 76%. As a result, the BERT model was selected as the final classifier.

Table 3. Seven kinds of text classification algorithms and their classification accuracy

Model	*Accuracy*
FastText	75.54%
TextCNN	73.24%
TextRNN	68.05%
TextRNN-Attention	73.07%
TextRCNN	74.89%
DPCNN	69.91%
Transformer	73.82%
BERT	81.71%

After completing the classification, we extracted 100 results from each category of the final results to test the usability of the constructed classifier. The recall rate and precision rate of each classification exceeded 80%, which meets the accuracy requirements and proves the reliability of the classification results.

Table 4. Sampling results of the classifier constructed by the BERT

Classification	*Recall*	*Precision*
Not updates	0.88	0.93
Functional updates	0.86	0.91
Business updates	0.97	0.99
Visual updates	0.84	0.98
Visual improvements	0.96	0.81
Existing functions improvements	0.93	0.86
Systematical improvements	0.93	0.97
Bug Fixes	0.97	0.99
Adaptability adjustment for iOS system	0.99	0.97
New content and activities	0.92	0.89

Step 3: Measuring the innovation output of apps

Every update point is measured on the nine categories with the classification algorithm (i.e., BERT). For example, if a specific update point represents the content of functional updates, the functional update dimension is measured as one. Next, this value considers the radical innovation dimension. Each app is measured in the two dimensions of incremental innovation and radical innovation by aggregating their values on the nine categories.

Findings

In the section, we combined the app types, innovation types, and degree of innovation with various visualization methods to demonstrate and explore the app's lack of innovation in depth.

Gradient distribution of the scores of app innovation

As previously indicated, app innovation output has obvious polarization. Nearly 10% of apps scored 0 in radical innovation in 2019. To the contrary, 35 apps performed very well in this aspect, with a score of over 100. To present this phenomenon more intuitively, we clustered the evaluation results of innovation through the KMeans algorithm. After 13 iterations, the final clustering result is shown in Figure 1.

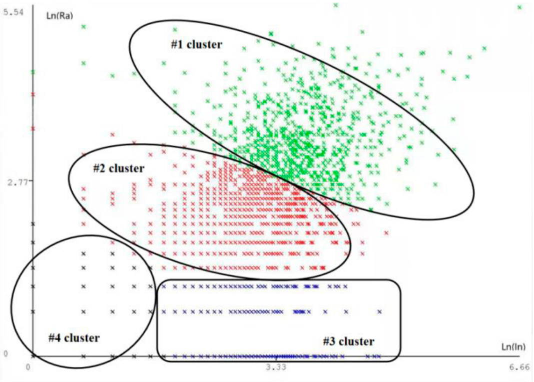


Figure 1. Clustering results of app innovation performance

As seen in Figure 1, the innovation performance of apps shows a gradient distribution. The #1 cluster accounts for 30% of the whole. Over 90% of apps in this cluster have excellent performance in both incremental innovation and radical innovation. They are most likely to be in the exploratory or growth period in the life cycle of software. For apps in the growth stage, speedy acquisition of new users through radical updates is the top priority. Meanwhile, it is necessary to quickly release incrementally updated versions based on user feedback to increase user retention. The #2 cluster is also the gathering place for killer apps such as Alipay and Taobao. These apps already have a stable stock of users, and there is room for growth in the number of users. The #3 cluster only consists of 19% apps; however, their characteristics are clear at a glance. They concentrate on developing incremental innovation, in most cases, because the users' groups are basically fixed, or they have technological innovations that are difficult for competitors to imitate. Apps in the #4 cluster are more cautious about innovation. Relatively speaking, these apps were rarely updated in 2019 and almost never ranked high in the IOS Hotlist. Nevertheless, WeChat, which is the most well-received instant messaging app in China, surprisingly belongs to the #4 cluster. There is no doubt that WeChat has become an essential app for almost everyone in China. For such a unique app, the risk of innovating (e.g., increase of input costs, users' dissatisfaction with the update) probably far outweighs the benefits.

Types and innovation modes of apps

Apps in the same field tend to imitate the successful ones to maintain functional equivalence (Wang, Li, & Singh, 2018), and their update modes usually maintain a certain degree of consistency in the sample of this study. Therefore, we calculated the average value of innovation performance for each type of app to observe their update patterns.

According to Figure 2, the radical innovation scores of the three types of apps, shopping, travel, and finance, are among the top three. As for the radical innovation dimension, game apps outshine others. By contrast, their radical innovation is mediocre. The scores of newspaper and magazine apps in these two dimensions are the lowest, but there are only four of them in our sample. Thus, it does not rule out the result of sample bias.

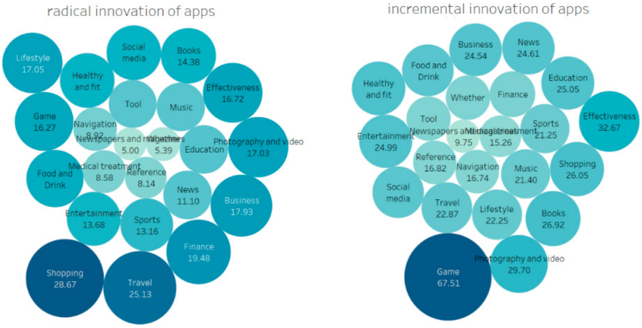


Figure 2. The average value of the scores of various apps in the radical innovation (left) and incremental innovation (right) dimensions

The game industry has always been an important branch of the mobile application industry. The majority of global app downloads in 2017 came from non-game apps. However, in terms of revenue, whether it is iOS or Android, game apps accounted for more than three-fourths of the earnings, and they are always the money-spinners.[1] In the iOS app store, games even have a special column.

The most obvious feature of game apps compared to other apps is that the new versions contain a lot of update points. For example, in the war strategy mobile game, The Clash of Kings, about 20 update points will be added for each update. Among the updates, it is not difficult to see that incremental innovation in content and activities is most essential, accounting for over 40% of daily updates. Holding periodic seasonal events and adding new characters, costumes, and props are the main ways for mobile games to keep users enthusiastic and make profits, rather than business updates. In addition, the properties of new characters and props are often modified based on user feedback, which is the primary content of existing function improvements.

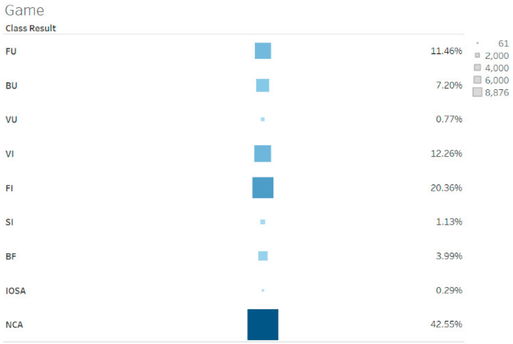


Figure 3. The proportion of various types of game app updates

[1] https://www.meetsocial.cn/information/78.html

Another interesting finding is that the update patterns of shopping, travel, and finance apps are extremely similar as shown in Figure 4. All of them did well in the dimension of radical innovation. Improvements to existing features are the dominating way they innovate incrementally, which accounts for approximately 20%. In addition, they also take notice of visual innovation and improvement. Compared with mobile games, user familiarity with the interface operation is not necessarily required. Therefore, they are more likely to revamp the UI to enhance the users' feeling of freshness and to better showcase the products or service. Among the three, shopping apps are most likely to think highly of commercial updates. Certainly, this is associated with the nature of the transaction of shopping. Travel and finance apps offer more digital content and functional service.

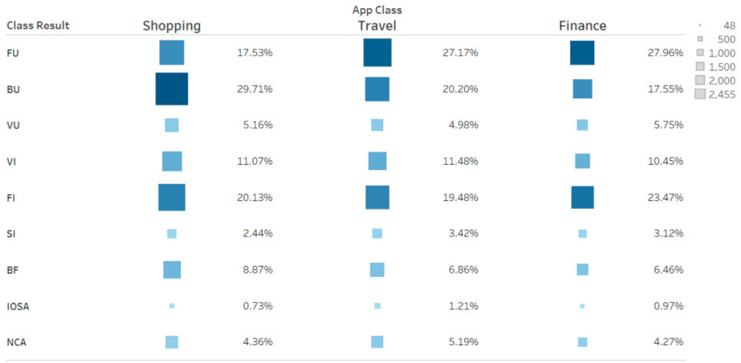


Figure 4. Proportion of various types of shopping (left), travel (middle), finance (right) app updates

Conclusion

To summarize, we have developed a measurement method for the innovation output of mobile apps, which can analyze tens of thousands (or even more) update documents and accurately identify the degree and type of updates. In addition, we applied this approach to conduct empirical research on 2,645 Chinese apps and found that the performance of the innovation output of apps is mainly divided into four levels. We also conducted an in-depth analysis of the update mode of some categories of apps.

Acknowledgements

Thanks are due to Xiaoke Zhang and Zhewen Liang for their great assistance with gathering the data of the apps. The funding from the National Natural Science Foundation of China under Grant 72072087 is also gratefully acknowledged.

References

Boer, H., & During, W. E. (2001). Innovation, what innovation? A comparison between product, process and organisational innovation. *International Journal of Technology Management, 22*(1-3), 83-107.

Christensen, C. M. (2013). *The innovator's dilemma: when new technologies cause great firms to fail*: Harvard Business Review Press.

Comino, S., Manenti, F. M., & Mariuzzo, F. (2019). Updates management in mobile applications: iTunes versus Google Play. *Journal of Economics & Management Strategy, 28*(3), 392-419.

De Brentani, U. (2001). Innovative versus incremental new business services: Different keys for achieving success. *Journal of Product Innovation Management: An International Publication of the Product Development & Management Association, 18*(3), 169-187.

Dewangan, V., & Godse, M. (2014). Towards a holistic enterprise innovation performance measurement system. *Technovation, 34*(9), 536-545.

Dziallas, M., & Blind, K. (2019). Innovation indicators throughout the innovation process: An extensive literature analysis. *Technovation, 80*, 3-29.

Edison, H., Bin Ali, N., & Torkar, R. (2013). Towards innovation measurement in the software industry. *Journal of Systems and Software, 86*(5), 1390-1407.

Evangelista, R., Sandven, T., Sirilli, G., & Smith, K. (1998). Measuring innovation in European industry. *International Journal of the Economics of Business, 5*(3), 311-333.

Fleischmann, M., Amirpur, M., Grupp, T., Benlian, A., & Hess, T. (2016). The role of software updates in information systems continuance—An experimental study from a user perspective. *Decision Support Systems, 83*, 83-96.

Geiger, S. W., & Cashen, L. H. (2002). A multidimensional examination of slack and its impact on innovation. *Journal of Managerial issues*, 68-84.

George, L. K., & Ferraro, K. F. (2016). Aging and the Social Sciences: Progress and Prospects - ScienceDirect. *Handbook of Aging and the Social Sciences (Eighth Edition)*, 3-22.

Hang, C., Neo, K., & Chai, K. (2006). *Discontinuous technological innovations: a review of its categorization.* Paper presented at the 2006 IEEE International Conference on Management of Innovation and Technology.

Kleinknecht, A., Van Montfort, K., & Brouwer, E. (2002). The non-trivial choice between innovation indicators. *Economics of Innovation and new technology, 11*(2), 109-121.

McIlroy, S., Ali, N., & Hassan, A. E. (2016). Fresh apps: an empirical study of frequently-updated mobile apps in the Google play store. *Empirical Software Engineering, 21*(3), 1346-1370.

Min Khoo, H., & Robey, D. (2007). Deciding to upgrade packaged software: a comparative case study of motives, contingencies and dependencies. *European Journal of Information Systems, 16*(5), 555-567.

Nelson, R. R. (2009). *An evolutionary theory of economic change*: harvard university press.

OECD, & Eurostat. (2018). *Oslo Manual 2018.*

Qiao, Z., Wang, G. A., Zhou, M., & Fan, W. (2018). The impact of customer reviews on product innovation: Empirical evidence in mobile apps. In *Analytics and Data Science* (pp. 95-110): Springer.

Soh, F., & Grover, V. (2020). Effect of Release Timing of App Innovations based on Mobile Platform Innovations. *Journal of Management Information Systems, 37*(4), 957-987.

Tian, H., Grover, V., Zhao, J., & Jiang, Y. (2020). The differential impact of types of app innovation on customer evaluation. *Information & Management, 57*(7), 103358.

Vaismoradi, M., & Snelgrove, S. (2019). *Theme in qualitative content analysis and thematic analysis.* Paper presented at the Forum Qualitative Sozialforschung/Forum: Qualitative Social Research.

Veryzer Jr, R. W. (1998). Discontinuous innovation and the new product development process. *Journal of Product Innovation Management: An International Publication of the Product Development & Management Association, 15*(4), 304-321.

Wang, Q., Li, B., & Singh, P. V. (2018). Copycats vs. original mobile apps: A machine learning copycat-detection method and empirical analysis. *Information Systems Research, 29*(2), 273-291.

Zhou, G., Song, P., & Wang, Q. (2018). Survival of the fittest: understanding the effectiveness of update speed in the ecosystem of software platforms. *Journal of Organizational Computing and Electronic Commerce, 28*(3), 234-251.

An Article-based Cross-disciplinary Study of Reference Literature for Indicator Improvement

Pei-Shan Chi[1] and Wolfgang Glänzel[2,3]

[1]*peishan.chi@kuleuven.be*
ECOOM, KU Leuven, Naamsestraat 61, 3000 Leuven (Belgium)

[2]*wolfgang.glanzel@kuleuven.be*
ECOOM & Dept. MSI, KU Leuven, Naamsestraat 61, 3000 Leuven (Belgium)

[3]*glanzw@iif.hu*
Library of the Hungarian Academy of Sciences, Dept. Science Policy and Scientometrics, Budapest (Hungary)

Abstract

In the nineties of the last century, researchers have applied several indicators to study reference literature of scientific articles. Glänzel & Schoepflin (1999) was the first time to capture and understand the subject characteristics in terms of structure and ageing of cited literature in the sciences and social sciences. Following and extending the pioneer study two decades ago, the present study focuses on how to build efficient instruments for the measurement of relevant aspects related to the 'hardness' of science. Apart from the observed general shift towards the use of more recent and indexed literature, the need of more than one single indicator is also detected in this study.

Introduction

Researchers' selection of references is influenced by social context which help reflect the sanctioned level by the community. These recognitions imply not only research track but also reference patterns specially associated with specific communication patterns across subject fields expressing their publication behaviour and co-authorship patterns, ageing of information in the mirror of diachronous and synchronous processes and as well as the "hardness" of science. The "hardness" of sciences revealing the degree of impersonality in the relationships of field members (Norman, 1967), is an indicator detecting different levels of publication patterns between the sciences and social sciences. The term "hardness" implies the strength of empirical evidence due to the laboratory experimental approach. Although its literal meaning may not be perfectly righteous, we remain this term in this study to reflect the original source and historical background for distinguishing the different patterns between the two types of sciences. Measures for the hardness degree based on the argument that "hard" sciences focusing on natural objects and phenomena while "soft" sciences focusing on human behaviours and activities (Norman, 1967) were proposed and applied by Price (1970), Moed (1989) and Fanelli & Glänzel (2013) in the context of references. According to Moed (1989, p. 474), the Price index which can help detect the hardness of sciences is "the percentage of references in the scientific literature to 0 to 4 years old other publications." Larivière, Archambault & Gingras (2008) found that Price Index values are decreasing since the mid-1950s and decreasing with the mean reference age. It is also confirmed on other studies that the Price Index is a decreasing function of the average and median reference age (Glänzel & Schoepflin, 1995; Egghe, 1997). Several studies reveal the characteristics of references and explore the information contained in references to detect borders between different sciences by the way scientists refer to their literature. For example, Glänzel & Schoepflin (1999) found that the percentage of references to serials is a sensitive measure of the communication behaviour between the sciences and the social sciences. It was reported in their study that about 80% of all science journals cite more than 70% of references to serials, while all social science journals refer to less than 70% references in serials. They also focused on differences in the ageing of journal literature in science and the social sciences. Interestingly, the characteristics of fields in the social sciences were found drifted towards those of the sciences, as Schubert & Maczelka (1993) pointed out that the research field of scientometrics had a clear move from those referencing characteristics of the ‘soft’ sciences towards those of the 'harder' sciences during the 80’s.

Given that the patterns of academic publishing reform during the past decades especially under the wave of Web 2.0, the characteristics of reference behaviour may be transformed as well and thus worth being traced after such a long period. In this study we will investigate reference patterns and the “hardness” of science measured by related indicators such as Price Index, reference age, and share of periodicals in references.

Furthermore, the granularity of the underlying subject delineation also plays a crucial part in the discriminative power of bibliometric indicators derived from publication and citation processes. Subjects fields defined with lower resolution tend to become more heterogeneous, the inflationary trends in scientists’ publication and citation behaviour as early noticed and discussed by Persson et al. (2006) further contributes to possible biases in bibliometric indicators at higher aggregation levels and requires proper normalization to keep their validity. These “inflationary trends” also foster the concentration on the most recent literature in the reference lists as being in line with literature growth and the intensification of network density (cf. Cozzens, 1985; Persson et al., 2004).

In addition, Moed (1989) concluded that “indicators reflect at least partly the size of the scientific activity in the subfield or topic”. That is why the measurement of research performance at lower aggregation levels could be distorted by indicators determined at higher levels of aggregation. This finally led him to breakdown the Price Index, usually calculated for the subfields or journals, to the level of individual papers.

In this context, it is also worthwhile to mention that a more recent study by Zhang & Glänzel (2017) already dealt with evolutionary aspects of literature ageing at the large scale. They compared the situation in 1992 (which by and large corresponds to the data used by Glänzel & Schoepflin, 1999) with that in 2014 and confirmed a growing trend in favour of citing older documents (cf. Verstak et al., 2014; Martin–Martin et al., 2016), which is partly due to Google Scholar search services and the digitalisation and electronic storage of scientific publications resulting in accessibility improvements to scholarly knowledge. Thus, the general decrease of the Price Index in all fields was accompanied by the growing mean age of references in practically all fields of the sciences including the social sciences.

In contrast to the study by Zhang and Glänzel, we base our present analysis on the article level to allow for the breakdown to subjects at any aggregation level and thus to avoid fuzziness caused, for instance, by field heterogeneity and journal multidisciplinarity.

The objective of the present study is threefold:

1. In how far have patterns changed in the time elapsed since the publication of the study by Glänzel & Schoepflin (1999)?
2. Do the findings regarding ageing-related and reference-structure indices by Glänzel and Schoepflin still hold for an efficient distinction between ‘softness’ and ‘hardness’?
3. Can we propose a new, multidimensional view on subject-related information use?

In order to avoid any effect of distortions caused by the particular choice of subject granularity, all indicators will strictly be calculated for individual papers and then aggregated within their disciplines. This makes it possible to apply the indicators also to multidisplinary journals, cognitively heterogeneous disciplines or even to use them in an interdisciplinary context.

Methodology

The hardness of a field can be measured by different indicators such as the Price index, the percentage of references to periodicals, the mean references age, the mean reference rate, relations among authors, epistemic networks, and the number of books in the subject borrowed from libraries (McGrath, 1978; Schubert & Maczelka, 1993; Wouters & Leydesdorff, 1994; Schoepflin & Glänzel, 2001). In this study the indicators applied to identify the hardness of publications are the mean reference count, the mean reference age, the percentage of WoS indexed references as a proxy for the share of references in serials, and the Price index. Only references with valid reference year and published after the year 1200 were analysed.

Data sources and data processing

The complete 2019 content of the WoS, including SCIE (Science Citation Index Expanded), SSCI (Social Sciences Citation Index), A&HCI (Arts and Humanities Citation Index), BKCI-S (Book Citation Index– Science), and BKCI-SSH (Book Citation Index– Social Sciences & Humanities) have been processed as source documents. Only the citable document type 'Article' and 'Review' which have more than 9 references each paper or book chapter were analysed in this study. In order to ensure compatibility and comparability with previous data as the possibility to look at evolutionary aspects we have restricted the selection of topics related with those used in previous studies. Fourteen out of 153 research areas[1] mapped by 254 Web of Science subject categories were selected to compare the reference patterns across four broad categories: Life Sciences & Biomedicine, Physical Sciences, Social Sciences, and Technology. The details of samples are listed in Table 1. The amount of book chapters in each area is insignificant to influence the result largely, therefore in this study the sample group is not further distinguished by publication type but includes all types of publication.

Table 1. Number of samples in the 14 analysed research areas

Subject Category	Research Area	Abbreviation	Nr. of items (Nr. of book chapters included)
Social Sciences	Archaeology	Arc	3,336 (150)
	Business & Economics	Bus	47,865 (2240)
	Psychology	Psy	50,843 (688)
	Sociology	Soc	8,212 (824)
Technology	Metallurgy & Metallurgical Engineering	Eng	20,493 (56)
	Information Science & Library Science	Inf	4,689 (166)
	Instruments & Instrumentation	Ins	21,406 (24)
	Transportation	Tra	8,658 (151)
Physical Sciences	Astronomy & Astrophysics	Ast	20,595 (144)
	Mathematics	Mat	62,410 (342)
	Polymer Science	Pol	24,167 (393)
Life Sciences & Biomedicine	Cardiovascular System & Cardiology	Car	29,904 (130)
	Immunology	Imm	26,470 (228)
	Research & Experimental Medicine	Med	31,278 (433)

Results

The results of all the indicators in selected research areas are shown in Table 2. The distinct characteristics can be identified across fields. For example, papers in research areas of the social sciences own relatively more reference counts, longer reference age, and lower Price index than other fields; by contrast, areas in life sciences have extremely high ratio of WoS indexed references. Figure 1 illustrates these differences across fields obviously. Areas in technology

[1] https://images.webofknowledge.com/images/help/WOS/hp_research_areas_easca.html

and life sciences & biomedicine have generally around 35% percent of references published within 5 years, while the Price indices of areas in social sciences are around 20% only. Mathematics is very different from other areas in physical sciences, owning a quite low ratio of recent references which is of the same magnitude as social sciences. In the social sciences, economics and psychology have distinct high ratios of WoS references as shown in Figure 1. They are closer to information science and mathematics, but still much less than those areas of medicine, engineering, physics, or chemistry.

Figure 2 shows that in general papers in technology field cite lowest numbers of reference, and papers in technology and life sciences have longer reference age. The latter tendency responds the patterns of Price index revealed in Figure 1. Most of areas show their reference behaviours in line with other areas of the same broader field, except for mathematics which is a unique subject citing very few references and most of them are relatively older and non-indexed publications. This may imply that mathematics is better to be excluded from the discussion of applying referencing behaviours to distinguish the field differences or the hardness of science.

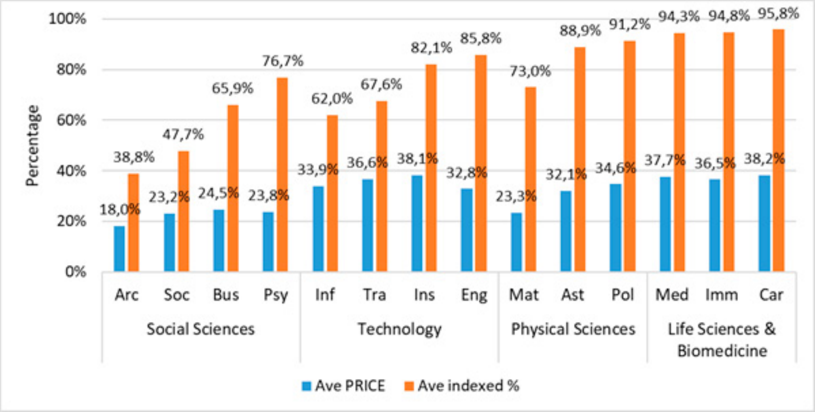


Figure 1. Average Price Index and average share of WoS references of papers published in 2019 in 14 research areas

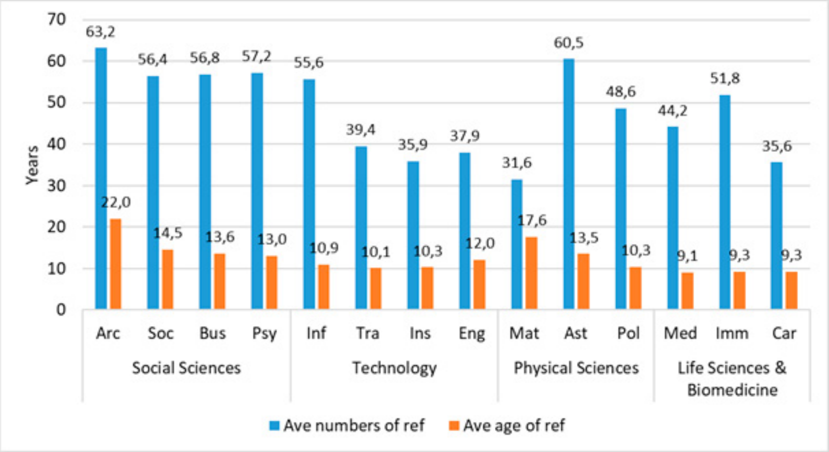


Figure 2. Average count and age of references of papers published in 2019 in 14 research areas

Glänzel & Schoepflin (1999) have used several indicators to study reference literature of scientific articles in the sciences and social sciences to capture and understand the subject characteristics in terms of structure and ageing of cited literature. In particular, they used the number of references, the share of references in serials, the mean reference age. The latter indicator was used to substitute the Price Index as the authors had found a strong (of course negative) correlation between these two indicators (Glänzel & Schoepflin, 1995). They concluded that the share of references in serials proved an efficient indicator of hardness with strong discriminative power. They also found a week correlation between the ageing and the reference-structure based indicators. The objective of the present study is not to redo the previous work by Glänzel and Schoepflin and to compare the results, or to simply add an evolutionary view, but to come to conclusions on how to build efficient instruments for the measurement of relevant aspects related to the 'hardness' of science. The strong correlations, on the one hand, and the almost uncorrelatedness between the considered indicators found by Glänzel and Schoepflin, on the other hand, already suggested that one single measure would certainly not suffice to reflect subject-characteristic reference patterns.

Table 2. Reference statistics of papers published in 2019 in 14 research areas

Broad category	Research areas	Nr. of samples	Mean nr. of refs	Mean Price Index	Mean % of WoS refs	Mean ref. age
Social Sciences	Arc	3,336	63.19	18.0%	38.8	21.98
	Bus	47,865	56.83	24.5%	65.9	13.59
	Psy	50,843	57.17	23.8%	76.7	13.05
	Soc	8,212	56.42	23.2%	47.7	14.51
Technology	Eng	20,493	37.94	32.8%	85.8	12.05
	Inf	4,689	55.63	33.9%	62.0	10.91
	Ins	21,406	35.94	38.1%	82.1	10.29
	Tra	8,658	39.45	36.6%	67.6	10.09
Physical Sciences	Ast	20,595	60.51	32.1%	88.9	13.52
	Mat	62,410	31.57	23.3%	73.0	17.59
	Pol	24,167	48.58	34.6%	91.2	10.35
Life Sciences & Biomedicine	Car	29,904	35.58	38.2%	95.8	9.30
	Imm	26,470	51.81	36.5%	94.8	9.27
	Med	31,278	44.16	37.7%	94.3	9.11

Table 2 gives the indicator values for 14 selected Web of Science research areas that have been chosen to connect results to those of the earlier Glänzel-Schoepflin studies. In all comparable research areas, we experience an increase of the reference age, which is in line with the observations by Zhang and Glänzel. Mathematics shows a strong growth in obsolescence (from 11.3 to 17.6, while the change remains rather limited with respect to the G&S study of 1999 (e.g., from 7.9 to 9.1 in research & experimental medicine, from 12.5 to 14.5 in sociology and from 9.1 to 10.9 in information & library science). This is accompanied by a growth of the share of database-indexed references, of course not typically for the life sciences, where the share was already close to or above 95% two decades ago. The comparison of the two indicators lets us already assume that both cannot clearly and most notably not simultaneously be associated with concepts like 'hardness'. The deviating values of the indicator pairs in all broad categories except the life sciences may illustrate that. Mathematics may serve as the most extreme example.

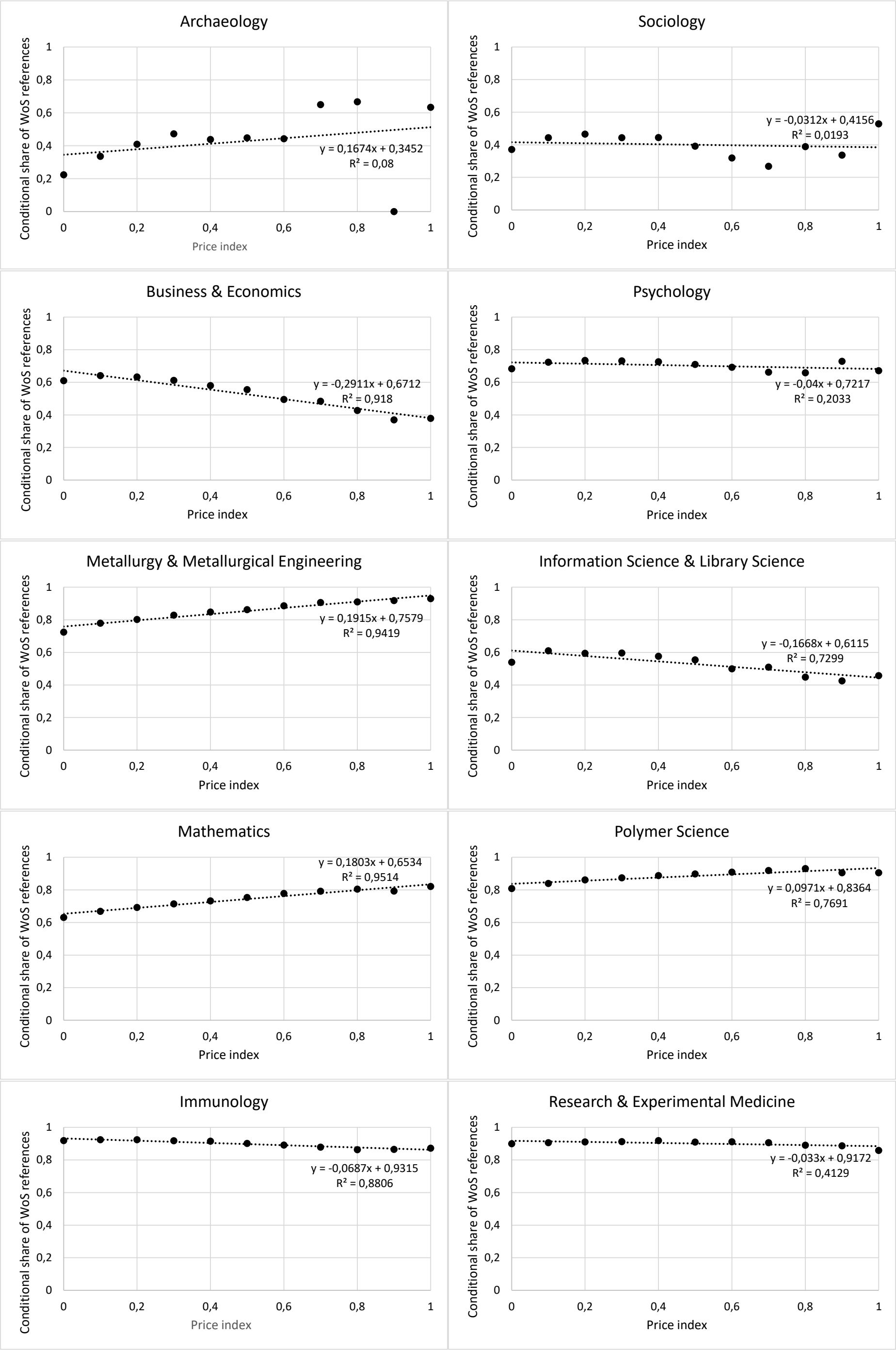


Figure 3. Scatter plots of conditional share of indexed references vs. Price Index in 10 selected research areas

The field is known for the slow ageing (researchers who have ever published or reviewed papers in this field understand why), but is, in turn, much relying on references to journals. Sociology with similar ageing characteristics originate information mainly from not indexed sources. In principle, none of the two indicators alone can express what we expect as a measure of hardness. We deepen this finding by regression analysis, showing the plots of the share of WoS-indexed references vs. Price Index in Figure 3.

We have chosen 10 out of the 14 research areas to ensure the legibility of charts, and calculated the conditional expectations based on the shares of WoS indexed and recent references (<5 years old), i.e., the Price Index. Instead of a direct linear regression analysis, we analyse another type of relationship (cf. Glänzel, 2010; Chi & Glänzel, 2018), namely by directly analysing the plot of π vs. the conditional expectation $E(\sigma|\pi)$, where σ represents the share of WoS indexed references and π stands for Price Index. This method has already successfully been applied by the authors in several contexts, among others for the correlation between usage and citation statistics (see Chi & Glänzel, 2018). There are two possible extreme cases: $E(\sigma|\pi) = E(\sigma) =$ constant means uncorrelatedness. The other extreme is the identity $\sigma = \pi$, with the trivial solution $E(\sigma|\pi) = \sigma$. However, because of the properties of conditional expectations there exists always a function f, such that $E(\sigma|\pi) = f(\sigma)$.

These plots are shown in Figure 3 where horizontal lines represent the extreme case of uncorrelatedness. In the selected fields in the life sciences, the two variables proved to be almost uncorrelated $E(\sigma|\pi) \approx \sigma$, where the slope is close to zero and the constant corresponds to the share of WoS indexed references. We just mention in passing that the most striking situation was found in research & experimental medicine, where the slope amounted to 0.005. In the life sciences, the natural sciences (except mathematics) and technology (except information science) the slopes are close to zero in general (also in the research areas not plotted here) and $E(\sigma)$ gives the expected share in these cases. This amounts to 0.9 or higher in the life sciences. A similar situation can be observed in the social sciences for sociology and psychology, where the slopes are close to zero but the $E(\sigma)$ amounts to 0.42and 0.72, respectively. We also mention that in instruments & instrumentation and transportation (not shown in Figure 4) with slopes close to zero, the expected shares of σ are 0.75 and 0.60, respectively.

By contrast, there are slight but distinct correlations between the two variables π and σ. These can be in part positive (e.g., for archaeology, metallurgy and mathematics) as well as negative (e.g., for business & economics and information & library science). In these cases, ageing as expressed by recentness of cited literature. Thus, for instance, in mathematics grows the share of WoS references with the share of recent references (and thus with assumingly decreasing age of sited literature), while in business & economics one can observe the opposite effect.

While, in the previous subsections, we have focused on selected research areas as defined by Clarivate Analytics, in this subsection we further lift up our perspective to observe the reference patterns across broad categories. The 14 research areas were applied to measure the general characteristics of their corresponding field. Figure 4, which gives the distribution of indicator values over individual papers in each major category, shows the basic statistics in terms of the two main indicators, share of indexed references and share of recent references (i.e. Price Index). In both graphs, the difference between life sciences and social sciences is obvious, especially for the share of indexed references. In the graph of Price Index, papers in technology have larger values than physical sciences and social sciences, although they have similar values of share of indexed references as physical science and much smaller values than life sciences. This finding emphasises the immediacy of references in the field of technology, and also, reveals the different function of the two indices distinguishing the major fields. For applied sciences such as technology, the Price index is more sensitive to detect its need of information immediacy, which is the most distinction from physical sciences. This finding suggests once again that the two indicators reflect essentially two different aspects of hardness.

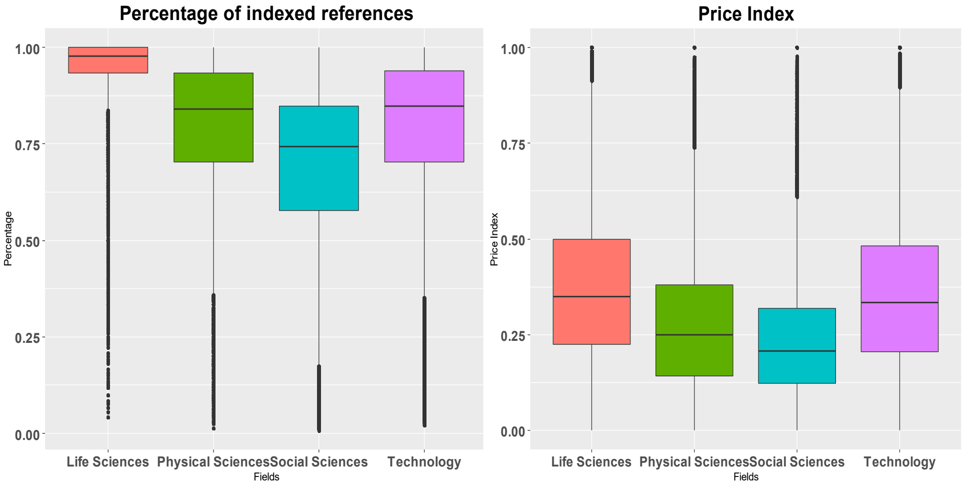


Figure 4. Box charts of share of indexed references and Price index in four broad categories

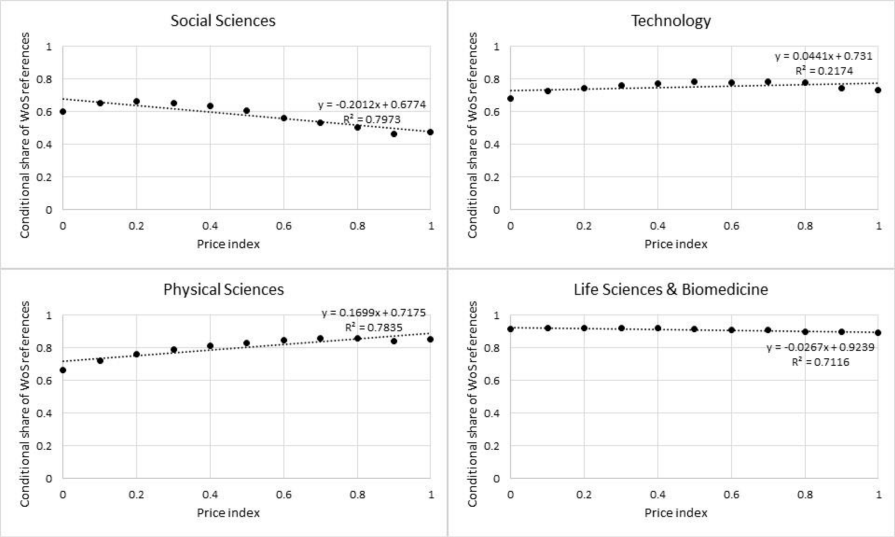


Figure 5. Scatter plots of conditional share of indexed references vs. Price Index in four broad categories

Following the analysis of the conditional expectations based on the shares of WoS indexed references and Price index, Figure 5 presents the plots in the four main fields. Similar to the findings based on individual research areas, mentioned in last section, we found a close-to-zero slop pattern in the life sciences, the natural sciences and technology. The life sciences own the highest expected share E(σ) (of 0.90), while technology and the natural sciences have similar share (of 0.70). Compared to these hard sciences, social sciences have the lowest share (of 0.67). In terms of the correlations between the two variables π and σ, social sciences and physical sciences have totally opposite consequent. The share of WoS indexed references grows with the share of recent references in physical sciences but decreases in the social sciences at the same ratio and condition.

Conclusions

This study, which was based on selected research areas and the four broad categories according to Clarivate Analytics Web of Science Core Collection, has brought three important answers to the questions addressed in the introduction. The evolution reflected by changing indicator values that could be observed is not unambiguous. There is certainly kind of shift towards "hardness" accompanied, even influenced but not necessarily caused by changing scholarly communication patterns. This shift is, on one hand, manifested by more frequently publishing in periodicals or serials with expectation of representing high standards, that is, preferably publishing documents that are indexed in the large multidisplinary bibliographic databases, and, on the other hand, by citing more recent literature, which is at least in part a (by-)effect of the global growth of literature (e.g., Cozzens, 1985; Persson et al., 2004; Larivière, Archambault & Gingras, 2008). A contrary effect certainly lies in electronic data storage and the improvement of search engines resulting in the accessibility of older and digitalised literature sources (see Verstak et al., 2014; Martin–Martin et al., 2016; Zhang & Glänzel, 2017). Consequently, the interpretation of the Price Index as the indicator of "hardness" of science, even in the context of ageing of scientific literature may need some adjustment. As to the second research question, the share of references in serials (or in indexed literature) being a structure-related indicator has considerably changed over time. In order to gain a more nuanced picture of "hardness", one would need more than one single indicator. The application of an indicator pair, one reflecting ageing, and the other one structural aspects, notably if built on individual documents, were able to capture the complex phenomena sketched by the above-mentioned studies and our present paper. This answers the third question.

References

Chi, PS. & Glänzel, W. (2018). Comparison of citation and usage indicators in research assessment in scientific disciplines and journals. *Scientometrics*, 116(1), 537–554.

Cozzens, S.E. (1985), Using the archive: Derek Price's theory of differences among the sciences. *Scientometrics*, 7(3-6), 431–441.

de Solla Price, D. J. (1970). *Citation measures of hard science, soft science, technology, and nonscience*. In C. E. Nelson, D. K. Pollock (Eds.), Communication among Scientists and Engineers (pp. 3–22), Lexington, MA, USA: Heath.

Egghe, L. (1997). Price Index and its relation to the mean and median reference age. Journal of The American Society for Information Science, 48(6), 564–573.

Fanelli, D., & Glänzel, W. (2013). Bibliometric Evidence for a Hierarchy of the Sciences. *PLoS ONE*, 8(6), e66938.

Glänzel, W., & Schoepflin, U. (1994). A stochastic model for the ageing of scientific literature. Scientometrics, 30(1), 49–64.

Glänzel, W. and Schoepflin, U. (1995). *A bibliometric ageing study based on serial and non-serial reference literature in the sciences.* Proceedings of 5th International Conference on Scientometrics and Informetrics, held in River Forest, IL, June 7–10. Learned Information, Medford, pp. 177–185.

Glänzel, W. & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing and Management*, 35(3), 31–44.

Glänzel, W. (2010). On reliability and robustness of scientometrics indicators based on stochastic models. An evidence-based opinion paper. *Journal of Informetrics*, 4(3), 313–319.

Glänzel, W. & Thijs, B. (2012). Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics*, 91(2), 399-416.

Glänzel, W. & Chi, P.-S. (2019). *Research beyond scholarly communication – The big challenge of scientometrics 2.0.* In G. Catalano, C. Daraio, M. Gregori, H. Moed, G. Ruocco (eds.), Proceedings of the ISSI Conference 2019, Rome, Italy, 424–436.

Larivière, V., Archambault, E., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: from exponential growth to steady state science (1900–2004). *Journal of the American Society for Information Science and Technology*, 59(2), 288–296.

Martin-Martin, A., Orduna-Malea, E., Ayllon, J. M., Lopez-Cozar, E.D. (2016). Back to the past: On the shoulders of an academic search engine giant. *Scientometrics*, 107(3), 1477–1487.

Moed, H. F. (1989). Bibliometric measurement of research performance and Price's theory of differences among the sciences. *Scientometrics*, 15(5), 473–483.

Norman, S. (1967). The hard sciences and the soft: some sociological observations. *Bulletin of the Medical Library Association*, 55, 75–84.

Persson, O., Glanzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, 60(3), 421–432.

Schubert, A. & Maczelka, H. (1993). Cognitive changes in Scientometrics during the 1980s, as reflected by the reference patterns of its core journal. *Social Studies of Science*, 23(3), 571–582.

Verstak, A., Acharya, A., Suzuki, H., Henderson, S., Iakhiaev, M., Chiung, C., Lin, Y., Shetty, N. (2014). *On the shoulders of giants: The growing impact of older articles*. http://arxiv.org/abs/1411.0275. Accessed 12 February 2021.

Wouters, P. & Leydesdorff, L. (1994). Has Price’s dream come true: Is Scientometrics a hard science? *Scientometrics*, 31(2), 193–222.

Zhang, L & Glänzel, W. (2017). A citation-based cross-disciplinary study on literature aging: Part I –the synchronous approach. *Scientometrics*, 111(3), 1573–1589.

Patterns of Knowledge Diffusion via Research Collaboration on a Global Level

Zaida Chinchilla-Rodríguez [1,2], Jialin Liu[2] and Yi Bu[2]

[1]*zaida.chinchilla@csic.es*
Instituto de Políticas y Bienes Públicos (IPP), Consejo Superior de Investigaciones Científicas (CSIC) (Spain)

[2] *{liu_jialin, buyi}@pku.edu.cn*
Department of Information Management, Peking University (China)

Abstract

The evolution of the patterns of collaborative strategies plays an important role in the social construction of science to design efficient research policies and to support the production of knowledge. As international research landscape is clearly undergoing continued structural change, we aim to analyze dynamics in collaboration strategies/patterns at the global and disciplinary level, and to investigate whether countries are collaborating more diversely within their region compared to outside their own region using different approaches. We found that internationalization in science is growing in all disciplines. The number of countries in the global network establishes stronger partnerships, forming tightly knit groups and spreading influence more widely among countries with a preference to collaborate within the region instead of collaborating with out-region countries. We believe that these results, though preliminary, have significant implications for the design of programmes and alliances.

Introduction

Science is being characterized by the exponential growth in publications and the rise of team science scholars. This trend of growth in the size of research teams have an effect on evaluative assessment, which are becoming more significant in all subject fields (Bozeman & Boardman 2014; Wuchty, Jones, & Uzzi, 2007). Individual and institutional research assessments supported by metrics have a powerful "pull" effect on scientists competing intensely within an economy of reputation seeking to maximize their own research output and impact (Hennemann & Liefner, 2015). At the same time the evolution of the patterns of collaborative strategies, especially international collaboration, plays an important role in the social construction of science to design efficient research policies and to support the production of knowledge (Coccia & Wang, 2016). The availability of local skills and capacities are central to the development trajectories of regions incorporating science, technology, and innovation in their national development agendas for economic growth (UNESCO, 2015).

It has been argued that "the best science comes from international collaboration" and "scientific research is entering a new age, driven by international collaborations" (Adams, 2013, p. 557). As scientific collaboration networks are expanding and characterizing the social construction of science (Adams, 2012), the growth of international collaboration has been studied to investigate collaboration at different levels and to inform scientific policy makers who evaluate the scientific output of their countries. Scientific policy makers in developed countries are definitely aware of the importance and benefits of research collaboration, with many countries providing specific funding to support large-scale collaborative research projects (Corley, Boardman, & Bozeman, 2006). The study of collaborative strategies in science might provide detailed patterns and inform science policy makers. However, little has been known about how collaborative strategies at the global and disciplinary level have evolved in the last decade.

The growth of international collaboration has been analyzed at the global, regional, and disciplinary level in previous studies. Wagner and Leydesdorff (2005), for instance, found that the principle of the preferential attachment might explain that countries with more collaboration are able to attract more collaboration. Indeed, a few countries dominate the knowledge flows at the global level, even when the global network has widened to include a much greater number of countries (Leydesdorff et al., 2013). At the regional level, the growth of international collaboration moves science system away from institutional collaboration, with stable and strong national collaboration. For example, the growth of publications in major European systems is almost entirely attributable to internationally co-authored papers. This growth is not only a function of state and EU promotion and funding but also reflects individual scientist's pursuit of reputation and resources (Kwiek, 2020). The globalization of science does not seem to have evolved uniformly across all countries and regions, as historical, geographical, and economic factors play a key role (Chinchilla et al., 2019; Geuna, 2015; Sherngell, 2013). Knowledge and capabilities have been conceptualized as geographically sticky, since tacit knowledge and abilities are a result of a workforce and not easily transportable. Research collaboration with scientists from other regions could allow countries to upgrade their academic capacity and respond to unique societal and economic challenges more readily (Fitzgerald et al., 2021; Frenken & Boschma, 2007; Hidalgo et al., 2007).

Yet, this existing body of research usually focuses on more advanced countries−many countries fall outside this taxonomy and are therefore understudied. To better understand the dynamics of the global system of science, we must move to a more comprehensive analysis. As international research landscape is clearly undergoing continued structural change with new leaders (in term of size) emerging from all corners of the globe (Fitzgerald et al., 2021), we aim to analyze how this shift in the geographical spread and economic capacities of countries has shaped a reconfiguration of the global collaboration network. Besides, as it is supposed that the corresponding authorship has a distinctive value in the byline of publications (Bu et al., 2019; Huang et al., 2011), we want to study different roles in the integration of new countries in the global network, namely leading or supporting collaborators' role.

Research objectives

There are two research objectives of this paper: On the one hand, we aim to identify dynamics in collaboration strategies/patterns at the global level and disciplinary level; on the other hand, we expect to investigate whether countries are collaborating more diversely within their region compared to outside their own region.

Data and methods

Data were retrieved from the in-house version of the Web of Science (WoS) maintained at CWTS of Leiden University for articles, reviews, and letters. Each publication in the dataset is clustered as five disciplines according to the algorithm of Waltman and Van Eck (2012). These disciplines are SSH (Social sciences and humanities), BIO (Biomedical and health sciences), PHY (Physical sciences and engineering), LIF (Life and earth sciences), and MAT (Mathematics and computer science). National/Regional-level collaborations were collected from 12,187,856 distinct publications retrieved for the

2008-2017 period, yielding connections among 216 countries[1]. Among these documents, 3,114,463 were internationally co-authored; this is 25.5% of the total. Document metadata was used to extract author affiliation data for each document. These data allowed for the construction of networks on the bases of collaboration (i.e., co-authorship between countries on a given document). We study three overlapping temporal periods (2008-2011, 2011-2014, and 2014-2017, respectively) to avoid fluctuations in countries with a very limited amount of data records.

We enrich bibliographical metadata by adding information of research and development expenditure (% of GDP)[2] and researchers in R&D (per million people)[3]. As expenditures take time in translating into outcomes, infrastructures, workforce, etc., we calculate the mean average of the previous 4-year-long period for each period under analysis. For example, for the period 2008-2011, we consider expenditures in 2005-2008, and for the period 2014-2017, we consider expenditures in 2011-2014. We analyzed affiliated document using full counting methods for all papers and for leading papers. For each country analyzed, papers were grouped into three mutually exclusive categories, based on the institutional addresses of the authors, named collaborative strategies: 1) papers that only have a single institution (no inter-institutional collaboration), 2) papers that have at least two institutions from the same country (domestic collaboration), and 3) papers that have at least two institutions from at least two different countries (international collaboration).

For the constructed collaboration network, we are particularly interested in these measurements:

- The number of edges: It denotes the number of ties between countries.
- Network density: It allows us to determine the degree of cohesion that exists among the nodes, revealing whether the network has a thick or thin consistency and the extent of connectivity among nations (Wasserman & Faust, 1999).
- The average degree of nodes: It measures the spread of influence across the networks (Hanneman & Riddle, 2005).
- Size of k-core: A k-core is the maximal subgraph in which each node has at least k connections to other nodes in the subgraph, despite how many links we have outside the subgraph. With this method, the densely connected area can be identified and in order to be included in the k-core, a node must have at least k links to other nodes in the k-core, regardless of how many other nodes they are connected to outside the k-core (Kong et al., 2019).

Besides, we also expect to examine whether countries are collaborating more diversely within their region compared to outside their own region. To this end, we employ the Shannon's entropy by regions. Specifically, we follow Fitzgerald et al. (2021) and calculate three measurements:

- Within-region entropy: $CE_{in}(i,t) = -\frac{1}{log(N^{(t)}-1)}\sum_{j\in J_u}\frac{n_{ij}^{(t)}}{\sum_{j\neq i}n_{ij}^{(t)}}\cdot log(\frac{n_{ij}^{(t)}}{\sum_{j\neq i}n_{ij}^{(t)}})$, where $N^{(t)}$ represents the number of countries in our dataset in the given period t,

[1] Some of the 216 countries are not real "countries". Rather, some are regions/territories. Yet, for the sake of simplicity, we use "country" in the following sections in this paper.

[2] Gross domestic expenditures on research and development (R&D), expressed as a percent of GDP. They include both capital and current expenditures in the four main sectors: Business enterprise, Government, Higher education, and Private non-profit. R&D covers basic research, applied research, and experimental development.

[3] The number of researchers engaged in R&D, expressed as per million. Researchers are professionals who conduct research and improve or develop concepts, theories, model techniques instrumentation, software of operational methods. R&D covers basic research, applied research, and experimental development.

$n_{ij}^{(t)}$ the number of collaborations between countries i and j in time period t, J_u the set of countries in the region of country i.

- Outside-region entropy: $CE_{out}(i,t) = -\frac{1}{log(N^{(t)}-1)} \sum_{j \in J_o} \frac{n_{ij}^{(t)}}{\sum_{j \neq i} n_{ij}^{(t)}} \cdot log(\frac{n_{ij}^{(t)}}{\sum_{j \neq i} n_{ij}^{(t)}})$, where J_o the set of countries outside the region of country i.
- The proportion of within-region entropy (as a measurement to understand the dynamics of inter- and intra-region collaboration diversity): $C(i,t) = \frac{CE_{in}(i,t)}{CE_{in}(i,t)+CE_{out}(i,t)}$.

We also expect to see how the expenditures on R&D of a country affects its performance on international collaboration and within-region collaboration. To this end, we implement two distinct regression models. In the models, we adopt the proportion of internationally collaborative publications and the proportion of within-region Shannon's entropy as the dependent variable, respectively. For the independent variables, we consider the average R&D investment proportion (among all its GDP between 2008 and 2017) and the number of researchers per one million population in the country. For both models, we use the income level of countries as a control variable in both models.

Yet, we found that the distribution of within-region Shannon's entropy is highly skewed. To this end, we cannot adopt the raw variable to a regular regression model. Instead, we use the percentile (rank) of this variable in the model in practice. Moreover, for a specific country, if any of the values of two independent variables is missing, we remove this country from our regression analyses. This results in 109 countries/regions remained.

Results

Overview

Table 1 provides an overview of data for the three periods. All indicators show a growth on international collaboration. The number of internationally collaborative papers has grown at a high rate (71.8%) than the total number of papers (53.4%). At individual level, the number of authors involved in internationally collaborative papers has risen by 81.3% much faster than the number of unique authors at the global level (66.55%). Besides, almost all countries have participated in the global international network as leading or supporting collaborators.

Table 1. Basic statistics of research collaboration by 4-year periods.

ALL	2008-2011	2011-2014	2014-2017	Growth rate
Number of papers	3846013	4879242	5898173	53.36
% international collaboration	23.6	25.2	26.5	12.05
% domestic collaboration	35.2	36.5	36.7	4.11
% single-authored papers	40.7	38.0	36.7	-9.96
Number of authors	6352677	8371230	10580467	66.55
% international collaboration	32.7	34.2	35.6	8.86
% domestic collaboration	50.7	51.8	52.4	3.37
% single-authored paper	46.6	43.5	40.1	-14.08
Number of countries	211	218	227	7.58
% of leading countries	98.1	96.3	95.6	-2.56

At the disciplinary level, Figure 1 shows that the percentage of internationally collaborative papers has increased around 12% in all disciplines. BIO and LIF present the highest values of growth in contrast with SSH that decrease 2.34%. In domestically collaborative papers, SSH also present a decline followed away from MAT (22.5% and 5.5%, respectively) in favor of increasing proportionally in single-authored papers (26.5% and 0.54%, respectively). The rest of disciplines present difference patterns, decreasing the percentage of single-authored papers in favor of international and domestic papers. Similar patterns are observed in the number of authors involved in distinct collaborative partnerships (Figure 2).

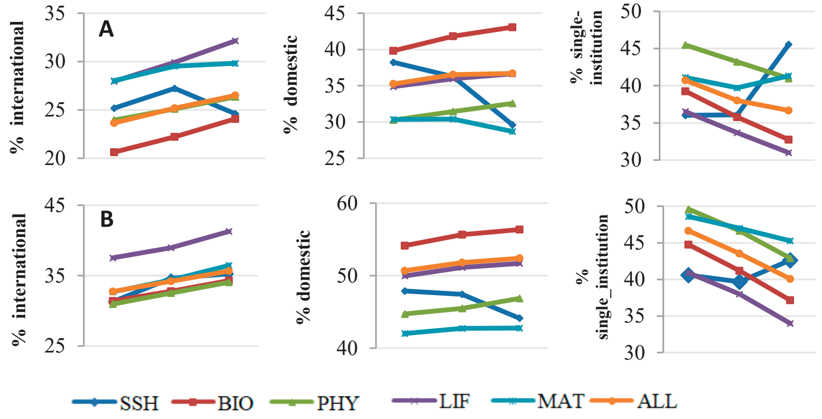


Figure 1. Evolution of the percentages of papers (A) and authors (B) by collaboration types at the global and disciplinary levels. Temporal periods: 2008-2011, 2011-2014, and 2014-2017.

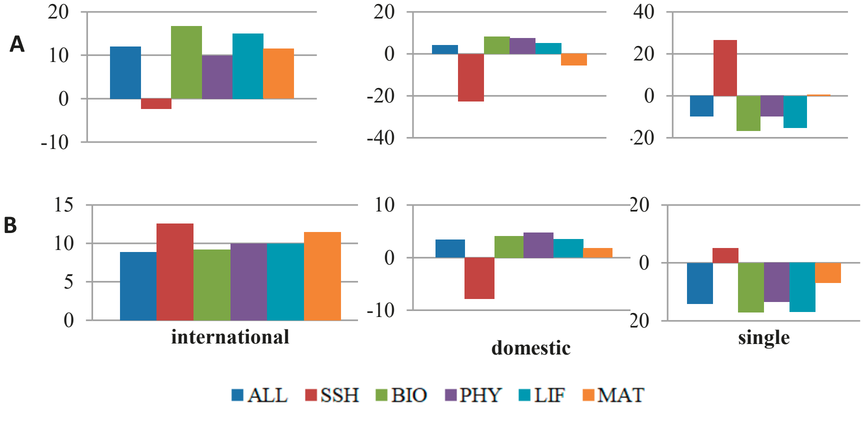


Figure 2. Growth rate of the number of papers (A) and authors (B) by collaborative types at the global and disciplinary levels.

The number of authors involved in at least one internationally collaborative paper increase significantly especially in SSH and MAT (12.6% and 11.5%, respectively). Considering authors that have been involved in at least one domestic paper, all disciplines show an increase except for SSH (-7.8%), which is the only discipline where the number of authors involved in at least one single-authored paper has risen (5%). Finally, considering the number of countries that at least have been involved in international collaboration as corresponding author, we can observe a descend especially strong in LIF and BIO and only PHY increase the number of countries acting as corresponding authors in international collaboration. These findings are in consonance with trends observed in previous studies (Coccia & Wang, 2017; Leydesdorff & Wagner 2008; Wagner, Park, & Leydesdorff, 2015).

Network analysis

We conducted network analyses to examine whether the network changed as the number of papers increased. Network measures are shown in two levels of aggregation: global and disciplinary level. Table 2 displays a set of basic structural indicators to analyze the structural properties of the global collaboration network. The number of countries linked to other countries has growth in the past decade (8%). According to the UNESCO[4], there are 193 country members and 11 Associate members. Our findings show that almost all nations and territories have researchers participating in international collaborative networks.

The size of the *k*-core component grew from 99 to 113 countries in a 10-year period. The average degree increases 34%. The average distance across the network is lower than two—a very low number considering the global network as a whole—and it is decreasing in the years examined. Diameter measures the longest distance between a pair of countries. In our study, the diameter is equal to 3 which mean that the distance from one side of the network to the other is small and the network is tightly linked together. Average clustering coefficient is high, which reinforces that the global network is forming tightly knit groups. To a certain extent, the global network is characterized by a high level of clustering and a small average number of steps between countries, which fits with the model of ‘small world (Watts & Strogatz 1998). All these indicators suggest that the network became denser, and that influence and power (in terms of collaboration flows control) were spread more widely among countries at the global level.

Table 2. Network statistics for the global network science.

	2008-2011	2011-2014	2014-2017	Growth rate
Number of nodes	210	217	227	8.10
Number of edges	8319	10031	12071	45.10
Density	0.38	0.43	0.47	24.14
Average degree	79.23	92.45	106.35	34.23
Average shortest path	1.63	1.58	1.54	-5.84
Diameter	3	3	3	0.00
Average clustering coefficient	0.79	0.80	0.81	2.83
Betweenness	0.003	0.003	0.002	-21.50
Size of *k*-core component	99	103	113	14.14

At the disciplinary level, Figure 1 shows the growth rate for each indicator over the period studied. We can observe that SSH registers the higher increase in the number of nodes (countries) in international collaboration followed by MAT and LIF. That means that these disciplines are becoming more opened to collaborative partnerships, especially in the case of SSH and MAT. By the other hand, the low increase of PHY and BIO might be explained by the cumulative/historic participation of science. Indeed, these fields are traditionally the most collaborative in research and they are characterized by a high participation of institutions around the world (Glänzel & Schubert 2004; Abramo,

[4] UNESCO https://en.unesco.org/countries/member-states

D'Angelo, & Murgia, 2013). In all indicators, we can observe a trend already discussed in previous research. The growth of internationally collaborative papers shows that new members (countries) are entering into the global network and ties among countries are increasing (edges). Greater density of the network and the decrease of betweenness measures suggest that fewer of the communications pass through the leading countries. According to Wagner et al. (2015), this may mean reducing influence for advanced countries, and shifting of power to some "peripheral countries".

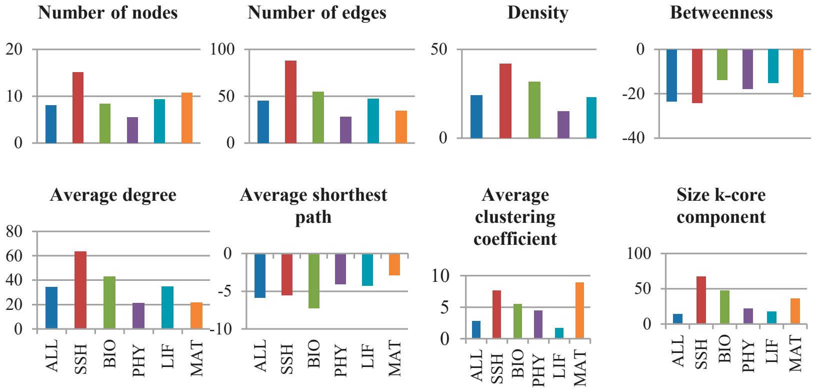


Figure 3. Network statistics by disciplines in the global network.

Collaboration diversity at the geographical level

So far, we have seen as the internationally collaborative papers are still growing in the last decade and that the network is becoming denser and more connected at the global and disciplinary level. The next step in our research is focused on regional level. We want to know how this internationalization is distributed at the geographical level, how it evolves over time and what the main strategies for building ties with partners inside or outside the region at the international level are.

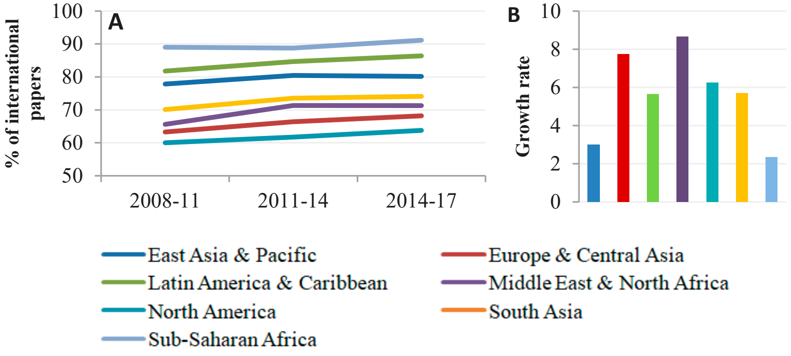


Figure 5. Percentage of papers by collaborative types and geographical regions.

Figure 5A shows the percentage of papers in international collaboration by regions. All regions tend to increase their international participation. Sub-Saharan Africa region shows the highest values in international collaboration followed by Latin America and

Caribbean. However, the regions that experienced the highest increases in their international presence are Middle East & North Africa and Europe (Figure 5B).

With country-level data in hand, it is possible to examine regional trends in global science over time to view whether large shifts occur in scientific output and knowledge absorptive capacity over geographic spaces (Wagner, 2018). The heterogeneity of countries in each group in terms of different levels of development and scientific production might affect these results. Then we expect to quantify how countries have changed (or not) their patterns of international collaboration measuring a countries' diversity of links to collaboration patterns both within their own region and with countries in other regions. As in Fitzgerald et al. (2021), we use the Shannon entropy of the distribution of collaboration partners for each country. The indicator provides values between zero and one. Values closer to one refer to countries that are collaborating evenly with many countries and values closer to zero refers to countries are collaborating with few countries. For each region, we plot the average proportion of within-region entropy (as a share of total entropy) for each region Figure 6. We observe that Europe & Central Asia shows the highest values of entropy, which means that countries from this region have increase their focus on diverse within-region collaboration. South Asia and North America present a lower diversity of partners. That might be explained by the number of countries in these regions. South Africa and North America are formed by 8 and 3 countries respectively in comparison with the rest of regions. Overall we observed a general declined over time of this tendency except for Sub-Saharan Africa. Countries from this region are intensely collaborating among them.

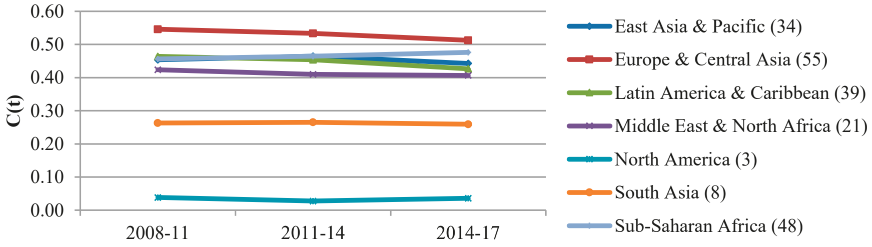


Figure 6. Within-region entropy (as a share of total entropy) results.

We plot collaboration entropy metrics within and outside of the region for each country by the final time period, with points scaled by the percentage of international collaboration. For example, in Figure 7, we can see that the South Asian countries strongly favor diverse partnerships outside their region; thus, they tend to collaborate more often with countries outside the region. Results show that almost all their production is carried out with international partners except for India, the country with the highest number of publications in the region, with the lowest percentages of internationally collaborative papers and the highest values of within-region collaboration diversity along with Maldives. We might suggest that growth of international publications in these countries is attributable to partnerships outside the region (Figure 7A). Figures 7B and 7C present the evolution over time of these two distinct strategies. We can see changes in some countries intended to opening up with their inter and intra-regional neighbors (for example, Maldives). On the other hand, there are countries that decreases intra-regional collaboration and either remain steady or decrease their inter-regional collaboration (for example, India and Pakistan respectively).

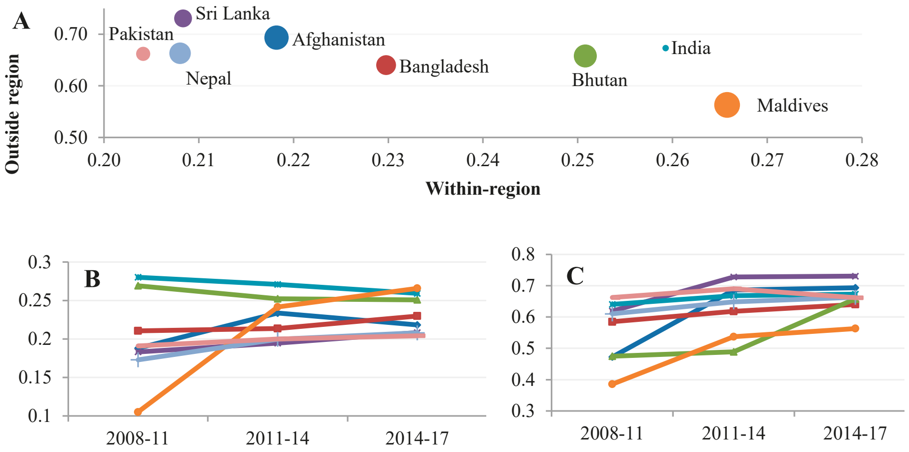


Figure 7. Inter and intra-region collaboration diversity for the South Asian countries (A); evolution over time of within-region collaborations (B) and outside region collaborations (C).

Regression analysis

Table 3 shows the regression result for the proportion of internationally collaborative publications. We see that when the number of researchers per million people and the income level is fixed, the proportion of internationally collaborative publications is negatively and significantly correlated with R&D expenditures. Yet, the number of researchers shows a positive relation with internationally collaborative paper count. Specifically, when other variables are equal, the proportion of internationally collaborative publications will increase 0.4% when there is one more researcher in every one million population.

Table 3. Regression results for the proportion of internationally collaborative publications (R Square=0.35, F=18.53, Sig.=0.00).

	Coefficient	Std.	t	P>|t|	[95% Conf. Interval]	
Intercept	98.767	5.307	18.611	0.000	88.245	109.290
R&D	-12.07	3.613	-3.341	0.001	-19.235	-4.907
Researchers	0.004	0.002	2.573	0.011	0.001	0.008

Note: "R&D" = Proportion R&D expenditures in the country's GDP. "Researchers" = Number of researchers per million population. The same below in Table 4.

Table 4. Regression results for the proportion of within-region entropy (R Square=0.05, F=4.84, Sig.=0.00).

	Coefficient	Std.	t	P>|t|	[95% Conf. Interval]	
Intercept	58.429	9.877	5.916	0.000	38.846	78.013
R&D	-13.434	6.724	-1.998	0.048	-26.767	-0.100
Researchers	0.011	0.003	3.370	0.001	0.004	0.017

Table 4 shows the regression model result for the within-region entropy. We observe that when other variables keep unchanged, the rank of within-region entropy for a country will decrease 12.3% (=13.434 / 109 [=number of countries in the regression model]) if the country increases 1% more GDP on R&D. Such an effect is significant on the 0.05 level. We also find a positive relation between the number of researchers per million population and the rank of within-region entropy for countries as well.

Discussion and conclusions

The evolution of the patterns of collaborative strategies, especially international collaboration, plays an important role in the social construction of science to design efficient research policies and to support the production of knowledge (Coccia & Wang, 2016). This study has shown the main collaborative strategies focusing on the growth of international collaboration over time at global, disciplinary and regional levels using different approaches. International collaboration in science is growing in all disciplines while the share of single-institutions papers is on the decline, which is in consonance with previous studies (Abt, 2007; Larivière et al., 2015). In terms of network analysis, the number of countries participating in the global network has growth establishing stronger partnerships, spreading influence and power more widely among countries and forming tightly knit groups. These findings are in accordance with previous studies showing that, between 1900 and 2005, the number of countries increased from 172 to 194, which represents a growth close to 13%. As long as new countries enter in the global network, this growth is decreasing and suggests that the network is not recreating political structures (Chinchilla et al., 2018; Wagner & Leydesdorff, 2005; Wagner, Park, & Leydesdorff, 2015). Our contribution here lies on the updating of temporal periods of international collaboration compared with other collaborative strategies.

Science policymakers care about the locations of national players as well as the outcomes of research, even when international collaboration seems to respond to the dynamics created by the self-interests of individual scientists rather than to other structural, institutional, or political factors (Wagner & Leydesdorff, 2005). In this study, we use the Shannon entropy as a way to determine the preference of countries to collaborate with other countries within or outside their regions showing that within-region collaboration has increased over time relative to international collaboration. We believe that these results, though preliminary, can shed light on the potential strategic programs or alliances. For example, in the cases that the diversity of collaborators increases more within regions than outside regions, a stronger regional clustering and more institutional cohesion inside the region should be observed. This is the case of Sub-Saharan countries where diversity within regions when compared to diversity between regions is increasingly higher, and the total strength of international collaborations relative to external collaborations also increases. It suggests the formation of localized regional collaboration networks. On the other hand, regions, e.g., Europe, with a high level of within-region partners seems to change the pattern over time, suggesting an opening up of global collaboration networks. However, the heterogeneity or not uniformity in scientific relationships of countries must be considered into each region

This internationalization presents policy challenges and opportunities. As internationalization of science affects the research performance of countries and, in certain degree, depends on the attractiveness of a partner in the global network, different strategies could be used to attract internal or external partners at convenience at different levels. At the individual level, the choice among forms of collaboration can be influenced by incentives and considerations concerning the organizational research system. For example, a scientist could be encouraged to privilege within-region collaborations because these will favor creation of team spirit within the workplace. In other cases, mechanisms for career advancement that reward external collaboration could encourage partnerships with regional and extra-regional colleagues (Abramo, D'Angelo, & Murgia, 2013; Acedo et al., 2006; Wagner & Leydesdorff, 2005). At the national and regional level, the forms and sources of research financing can influence the types of collaboration chosen. For example, it could encourage local- and intra-region- level collaborations, while supra-national financing and in certain cases, the incentive systems to individual organizations, can favor collaborations at the international level (Hoekman et al., 2010). A smart balance must be found between the necessary reinforcement of internal coherence, mainly around the project of collectively "discovering" the relevant specialization, and the challenge of positioning local research in the global world of science. In fact there is no contradiction if the territorial strategy is well defined (Benaim, Herauld, & Mérindol, 2016). In addition, it raises many interesting yet far-reaching questions, such as: which collaborators countries are drivers in fostering and enhancing smart specialization, clusters of innovation, absorptive capacities? Such questions deserve further investigation and suppose future research.

The current study has some limitations. First, scientific capacities of countries are driven by a combination of factors. We consider some of them, such as the income level, expenditures in R&D, and researchers. However, it is necessary to keep analyzing other factors (e.g., scientific capacities, power relationships) affecting international collaborations to better understand conditions and possibilities of the fewer developing countries to access to the global scientific network. Second, as scientific system of science is growing; the scale of bibliographical databases has grown, too, with new journals added in the collection. That means that comparisons over time might differ from current results. Third, in network analysis, the current paper does not consider whether peripheral countries are consolidating positions in the regional and global network. We plan to further study roles of countries (closeness, betweenness, etc.), in the regional and global network. Besides, as the vast differences in volume of production and scientific capacity by country complicate analyses of collaboration, we plan to analyze the temporal evolution of the exchange of knowledge between countries and the relative importance of specific countries in building ties across borders by applying similarity measures (e.g., Affinity Index, Probabilistic Affinity Index (Chinchilla-Rodríguez et al., 2021)) and combining regression analysis with other indicators. The ultimate goal is to help policy-makers in planning research agendas and collaborative alliances.

Acknowledgements

The authors would like to thank Rodrigo Costas for providing access to the CWTS in-house Web of Science database and work on the compilation of the dataset.

References

Abt, H.A. (2007). The future of single-authored papers. *Scientometrics*, 73(3), 353-358.

Abramo, G., D'Angelo C.A., Murgia G. (2013). The collaboration behaviors of scientists in Italy: A field level analysis. *Journal of Informetrics*, 7(2), 442–454.

Adams, J. (2012). Collaborations: The rise of research networks. *Nature, 490*(7420), 335–336.

Adams, J. (2013) Collaborations: The fourth age of research. *Nature, 497*(7451), 557–560.

Benaim, M., Heraud, JA., Mérindol, V. (2016) Scientific Connectivity of European Regions: Towards a Typology of Cooperative Schemes. Journal of Innovation Economics & Management, 3(21), 155-176.

Bu, Y., Zhang, C., Huang, Y., & Sugimoto, C. R., & Chinchilla-Rodríguez, Z. (2019). Investigating scientific collaboration through the sequence of authors in the publication bylines and the diversity of collaborators. In *Proceedings of the 17th International Conference on Scientometrics and Informetrics (ISSI 2019) (pp. 2300-2305)*, September 2-5, 2019, Rome, Italy.

Chinchilla-Rodríguez, Z., Miao, L., Murray, D., Robinson-García, N., Costas, R., & Sugimoto, C.R. (2018). A global comparison of scientific mobility and collaboration according to national scientific capacities. *Frontiers in Research Metrics and Analytics*, 3, 17.

Chinchilla-Rodríguez, Z., Sugimoto, C.R., & Larivière, V. (2019). Follow the leader: On the relationship between leadership and scholarly impact in international collaborations. *PLoS ONE, 14*(6), e0218309.

Chinchilla-Rodríguez, Z., Bu, Y., Robinson-Garcia, N., & Sugimoto, C. R. (2021). Examining the performance of different probabilistic affinity index (PAI) definitions in international scientific collaboration: A methodological exploration. *Scientometrics, 126*(2), 1775-1795.

Coccia, M., & Wang, L. (2016). Evolution and convergence of the patterns of international scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America, 113*(8), 2057-2061.

Corley, E.A., Boardman, P.C. & Bozeman, B. (2006). Design and the management of multi-institutional research collaborations: Theoretical implications from two case studies. *Research Policy, 35*(7), 975–993.

Fitzgerald, J., Ojanperä, S., & O'Clery, N. (2021) Is academia becoming more localized? The growth of regional knowledge networks within international research collaboration.

Hanneman, R.A. & Riddle, M. (2005). *Introduction to social network methods*. Riverside, CA: University of Riverside.

Hennemann, S., & Liefner, I. (2015). Global Science Collaboration. In *the Handbook of Global Science, Technology, and Innovation*, edited by D. Archibugi and A. Filippetti, 343–363. Somerset, NJ: Wiley.

Hidalgo, C.C., Klinger, B., Barabasi, A-L., and Hausmann, R. (2007). The product space conditions the development of nations. *Science, 317*(5837), 482-487.

Hoekman, J., Frenken, K., & Tijssen, R. J. W. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy, 39*(5), 662-673

Huang, M.-H., Lin, C.-S., & Chen, D.-Z. (2011). Counting methods, country rank changes, and counting inflation in the assessment of national research productivity and impact. *Journal of the American Society for Information Science and Technology, 62*(12), 2427–2436.

Frenken, K., & Boschma, R. A (2017). A theoretical framework for evolutionary economic geography: Industrial dynamics and urban growth as a branching process. *Journal of Economic Geography, 7*(5), 635-649.

Kwiek, M. (2020). What large-scale publication and citation data tell us about international research collaboration in Europe: changing national patterns in global contexts. *Studies in Higher Education*, 1-21.

Lariviere, V., Gingras, Y., Sugimoto, CR., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. Journal of the Association for Information Science and Technology, 66(7), 1323-1332. https://doi.org/10.1002/asi.23266.

Leydesdorff, L., & Wagner, C. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics, 2*(4), 317-325.

UNESCO. UNESCO Science Report: towards 2030. Technical report, Imprimerie Centrale, Luxembourg, 2015.

Wagner, C. S., Brahmakulam, I., Jackson, B., Wong, A., and Yoda, T. (2001). *Science and Technology Collaboration: Building Capacities in Developing Countries*. Santa Monica, CA: RAND

Wagner, C. S., Park, H. W., & Leydesdorff, L. (2015). The continuing growth of global cooperation networks in research: A conundrum for national governments. *PLoS One, 10*(7), e0131816.

Wagner, C. S. (2018). The collaborative era in science. Palgrave MacMillan, Chapter 7, p. 123, doi: 10.1007/978-3-319-94986-4.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378–2392.

Wasserman, S. & Faust, K. (1999). Social network analysis-methods and applications. Cambridge: Cambridge University Press.

World Bank. World Development Indicators. 2021. Available online: https://data.worldbank.org/topic/science-and-technology (accessed on 11 April 2021).

Countries' research priorities in relation to the Sustainable Development Goals

Hugo Confraria[1], Ed Noyons[2] and Tommaso Ciarli[3]

[1] *h.confraria@uece.iseg.ulisboa.pt*
UECE/REM- ISEG, Universidade de Lisboa, Portugal
SPRU, University of Sussex, United Kingdom

[2] *noyons@cwts.leidenuniv.nl*
CWTS, Leiden University, The Netherlands

[3] *ciarli@merit.unu.edu*
UNU-MERIT, Maastricht University, The Netherlands
SPRU, University of Sussex, United Kingdom

Abstract

We analyse the extent to which countries' research priorities align with their performance in the UN Sustainable Development Goals (SDGs). Our central assumption is that a misalignment between the investment in research areas and the socio-economic challenges may reduce the effectiveness of the investments in research to address those challenges. We develop a new method to identify research that is related to an SDG by examining which research areas in WoS have more publications that contain text that is related to SDG policy outlets. Then we propose a new method that combines all SDG indicators available, to measure the performance of countries in certain SDGs in relation to the top performers. Overall, we find that the SDGs' research priorities in which researchers in low and middle-income countries publish most in international journals, are not necessarily the research areas where those countries perform worst when analysing SDG indicators. SDG2 (Zero hunger) and SDG3 (Good health and well-being) are exceptions where we can find a degree of alignment, but in other challenges such as SDG12 (Responsible consumption and production) and SDG13 (Climate action), higher income countries perform worst and are not specialised in those research areas.

Motivation

Scientific research is a critical ingredient to develop knowledge-based economies, where knowledge drives productivity, social wellbeing and the achievement of socio-economic needs. Without scientific capacity, the skills and capabilities available in a country are constrained, and therefore, the ability to absorb, adapt and develop new ideas and technologies is limited.

Simultaneously, there is an increasing demand for science and research funding to be better aligned with socio-economic needs (Ciarli and Ràfols, 2019; Sarewitz and Pielke Jr., 2007). In this context, research should not be limited to frontier technologies, but the priority should be to contribute to solving our major collective challenges, that usually are more problematic in low- and middle-income contexts.

An important approach to understanding how countries and regions are making progress in solving societal challenges is to map efforts in relation to the targets and indicators set out in the seventeen 2030 Sustainable Development Goals (SDGs). These SDGs recognise that ending poverty and other deprivations must go hand-in-hand with strategies that improve health and education, reduce inequality, and spur economic growth – all while tackling climate change and working to preserve our oceans and forests.

However, from the science side, the process by which scientific endeavours influence achievements in SDGs (as well as other 'Grand' or societal challenges) is complex, distributed in space and time, and the links are usually indirect. Still, it is crucial to identify how certain research may be associated with individual objectives. Since funding resources are scarce and research funders need to make choices about the types of research they support, it is crucial to develop frameworks (as well as tools, datasets and methods) that allow for making informed choices and prioritising some research agendas over others.

In this paper, we use the SDGs to investigate the research priorities of the research systems in most countries in the World versus the main challenges they face. Our central assumption is that a misalignment between the investment in research areas and the socio-economic challenges, may reduce the effectiveness of the investments in research to address those goals, and we intend to bring prioritisation and directionality to the debate about science policy investment.

Results can help local policy makers and international donors in two crucial ways: first, to influence/support prioritisation in relatively neglected research areas, when such underinvestment may harm the achievement of socio-economic goals. Second, to have a better understanding (and more productive discussions with relevant stakeholders) of the research being prioritised, capabilities created in the country, and what reasons might have led to the absence of research in areas that are crucial to advance relevant socio-economic outcomes.

Data and Methods

We created an indicator of revealed research priorities of countries in a certain SDG[1] based on Web of Science (WoS) publication data, and an indicator of revealed socio-economic challenges of a particular country in a certain SDG based on a combination of indicators associated to the SDGs. Below we describe how we calculate the indicators that allow analysing these two dimensions.

Revealed research priorities by country in each SDG

Scientific publications were assigned to a specific SDG using a 2-step method developed by the authors. First, we build a query with a set of terms (or combinations of terms) that are strongly associated to a specific SDG. Second, we use those SDG related queries to search publications in WoS research areas that are generated by citation relations between all WoS publications in 2015-2019. Since these research areas are obtained from a publication-level clustering algorithm based on direct backwards and forward citations, the advantage of this approach is that these clusters can be seen as a community of papers that address (or are related to) a specific area/field. In comparison, a query-based approach is only able to identify individual publications. Our approach allows to include relevant publications that do not explicitly use a specific SDG query term in their abstract/title, because they use a slightly different language, or focus on an issue that was not explicitly mentioned in SDG policy reports and publications. For instance, with reference to health and well-being (SDG3), policy documents may refer to some tropical diseases, but not all of them. Our approach allows to expand the initial set of publications to other publications that are closely related. Since research is cumulative and the communities/groups of researchers are the ones able to solve a certain challenge (instead of a specific paper), we believe this approach is more adequate for our research question.

The methodology used to obtain a final list of keywords per SDG consists of a series of steps. In the first step, we selected texts from various sources that contained descriptions of specific SDGs. Instead of relying only on official UN sources to identify relevant terms, we chose to search a wide array of policy reports, grey literature, scientific publications, web forums and official UN sources. In this way, we aim to capture a broader understanding of SDGs that is shared in different types of publications and authors. We then extracted relevant fragments

[1] We decided to only include 16 (1 to 16) out of the 17 SDGs, because SDG17 deals with “Strengthening global partnerships and enhancing the means of implementing the Goals” which is very difficult to operationalise in terms of defining a query of terms strongly associated to that goal.

from these texts, which referred to a particular SDG and met a certain criterion.[2] This step allowed us to exclude text content that is not SDG specific and text that is about more than one SDG. Once selected, we partitioned the text referring to an SDG in different entries, respecting the authors' different ideas. Afterwards, we selected relevant keywords within these pieces of text using a combination of Textrank and Vosviewer algorithms. Finally, we carried out a manual selection of the keywords extracted through these filters and shared these lists with other team members and experts to check for missing or irrelevant terms.

After validating all 16 SDG queries, we applied those queries to the CWTS WoS dataset, in a sub selection of publications from 2015-2019, to search clusters of publications (research areas) strongly associated with each other by their backward and forward references. We use a classification system generated by Waltman and Van Eck (2012) at CWTS that separates all WoS in 4013 micro-clusters of publications. The algorithmically created clusters at this level prove to be an optimal granularity for normalisation (Ruiz & Waltman, 2015) and to be described by automatic labels and hence the preferred level to be used in our study. Then we run our 16 SDG queries in all publications of those clusters, to check which of the research areas have a high share of publications which include our SDG related text in their abstracts and titles. We then associated with a certain SDG all publications that belong to a cluster that has a relatively high share of publications with a certain SDG query term. Given that different SDGs required different parameters (shares), together with other team members and experts, we decided to define two thresholds for each SDG to allow for robustness checks later on:

1. Min threshold (Loose): the share of seed publications (containing the SDG keywords) over all publications in the research area that identifies a relevant number of research areas related to the SDG, among some non-relevant areas, based on information contained in the cluster labels[3] (micro and meso clusters) and in a sample of top publications (not containing the SDG keywords) against the SDGs targets.
2. Max threshold (Strict): the share of seed publications (containing the SDG keywords) over all publication in the research area that identifies only research areas that are strictly related to the SDG, based on information contained in the cluster labels (micro and meso clusters) and in a sample of top publications (not containing the SDG keywords) against the SDGs targets.

After this first initial analysis of what clusters are associated to what SDGs, we compared our results with the results obtained using the publicly available queries from the SIRIS approach.[4] By checking the differences between "loaded" clusters using the two approaches, and analysing the labels of those clusters, this comparison allowed us to improve our:

- Recall (type II error, false negative), namely clusters that should be associated with a certain SDG but are not with our current approach (queries).
- Precision (type I error, false positives), namely publications clusters that are currently being associated with a certain SDG but shouldn't.

After comparing our results using the two approaches and changing some of our terms to improve recall and precision, we recalculate the min-max thresholds and use the final set of

[2] These fragments must contain text referring specifically to at least one SDG. The text must refer to problems associated to the SDG(s), making a clear connection between the problems and Goal(s) (e.g. using the term "Sustainable Development Goal" in their analysis of the problem). References to "sustainability" alone were not considered sufficient for the document to be included. References to issues associated to the SDGs (e.g. poverty or hunger) but with no explicit mention of the SDGs were also not considered sufficient for the document to be included.

[3] The analysis of most frequent keywords from relevant clusters in both approaches was crucial to check differences and understand which terms retrieve which clusters.

[4] http://science4sdgs.sirisacademic.com/

research areas per SDG to do our country analysis. In this paper, results correspond to our Max thresholds only for simplicity. We ran a sensitivity analysis between Min and Max thresholds and the correlation of results between the two is extremely high. Each SDG specific thresholds can be provided upon request, and the platform[5] that we use to understand which publication researcher areas are associated with an SDG is openly available (Rafols et al., 2021)
The metrics for comparisons between countries are created based on address criteria, using fractional-counting method (counts are weighted by number of authors' countries and we only consider independent countries that have more than 500 publications in 2015-2019). Revealed research priorities of countries by SDG are calculated using a comparative specialisation index (Balassa, 1965) that allows to check if a country's research is more or less specialised in a certain SDG than the world average (Ciarli and Ràfols, 2019). Later we normalise this estimate of research specialisation between -1 and 1 (1 = High specialisation; -1 = Low specialisation) to be able to compare it with the revealed challenges.

Revealed challenges by country in each SDG

To analyse the performance of countries in different SDGs, we build an index per SDG that combines data from two different sources: i) UN SDG database; ii) SDG Index. We check which indicators have fewer missing values for all countries and years of interest and build a unique dataset with 80 different indicators. After compiling this dataset of indicators per SDG, we run a principal component analysis per SDG to obtain a single score by country/SDG for which data is available.[6] We followed the next steps:

- For the selected indicators, we calculate the logarithm of those that are not percentages or indexes (e.g. per capita)
- Then, we do a linear transformation, by converting each indicator/country to a score between 1 (Best) and 0 (Worst):

$$n_{cti} = \frac{Worst_{ti} - x_{cti}}{Worst_{ti} - Best_{ti}} \quad \text{c (country), t (period), i (indicator)}$$

- We reverse some variables for consistency, forcing higher values to represent better results.
- For each variable, we calculate the relative distance ($d_{cti} = p95_{ti} - n_{cti}$) of each indicator/country to the frontier of that indicator (top5% - percentile 95), and we changed all values below zero to zero. After this transformation, higher values represent worst results with respect to the SDG targets (higher challenges relative to countries at the frontier).
- We calculate z-scores for each relative distance to the frontier (top5%).

$$z_{cti} = \frac{d_{cti} - \mu}{\sigma} \quad \mu\text{(average)}, \sigma\text{(std. deviation)}$$

- We compute a principal component analysis (PCA) (Jackson, 1991) for each SDG with more than one indicator available, and we forced the PCA to estimate only one component per SDG (eigenvalues and eigenvectors can be provided upon request).
- We predict the scores of all SDGs for all countries and we normalised the results between -1 (High performance) and 1 (1 = Low performance).

[5] https://public.tableau.com/profile/ed.noyons#!/vizhome/UKStringsSDGtocommunities/Dashboard1

[6] In this version we present results for the period 2008-2017, which is when we are able to get more data points for most countries/years. In the conference we are planning to present results also for time dynamics between two separate periods (2008-2012 and 2013-2017).

Countries with low performance in a specific SDG can be seen as those furthest away from the "frontier" in that specific SDG, meaning that those countries have a higher socio-economic challenge in that SDG. Countries with high performance are those at the "frontier" that are the best performers in that specific SDG.

Preliminary results

We start by analysing descriptively how many publications per SDG exist in the World, according to our approach, and how certain income country groups perform a larger share of SDG research than others. In the second part, we graphically analyse the relation between SDG research priorities and SDG challenges for all countries for which data is available.
According to our approach (using both the strict and loose threshold), during 2015-2019, the research in WoS that is related to at least one SDG is between 30% (strict threshold) and 50% (loose threshold). The remaining research can be seen as research not directly related to the SDGs' targets, indicators and objectives.

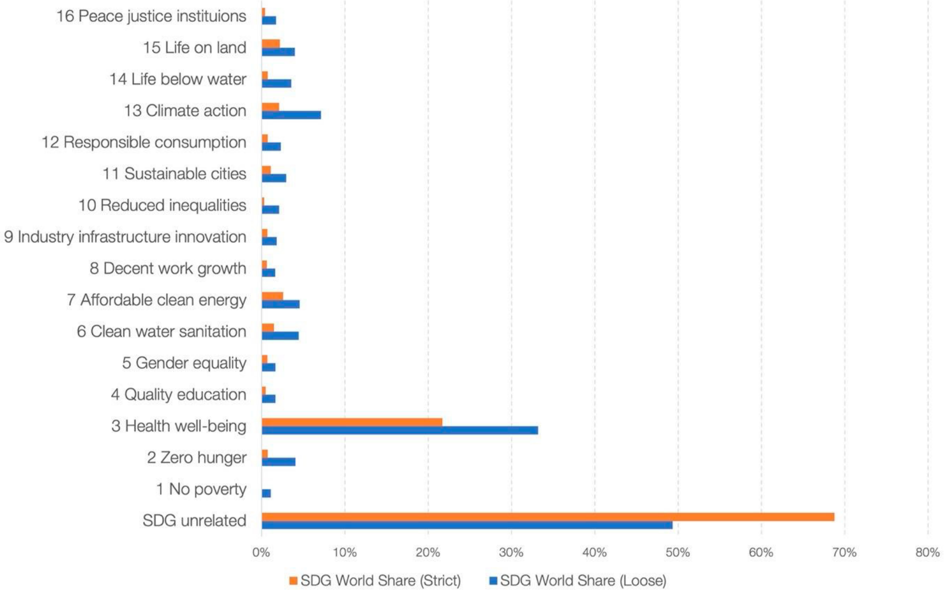


Figure 1. Share of publications associated to an SDG between 2015 and 2019 in the World.
Source: WoS

In Fig. 1, we observe that SDG3 (Good health and well-being) is the SDG with higher share of research associated to it. At the other spectrum, we find that SDG1 (No poverty) is the SDG with lower share of research associated to it. It is important to note that research related to a specific SDG can be about numerous issues. For example, in SDG3 (Good health and well-being) we find a lot of research related to cancer, cardiovascular diseases, infectious diseases but also some research related to social sciences (e.g. mental health) and ambient air pollution. Therefore, it might be the case that within a SDG there might be specific misalignments of research priorities versus societal challenges if, for example, most of the research in SDG3 is related to diseases that are not the ones that generate more burden globally (Confraria and Wang, 2020; Evans et al., 2014; Yegros et al., 2019). Another aspect worth noting in Fig.1 is results depend greatly on the threshold used. For example, the amount of research related to SDG1 while using the loose threshold is more than 15 times higher than while using the strict threshold. In SDGs 3 (Good health and well-being), 7 (Affordable and Clean Energy) and 15 (Life on Land) the differences are not that large (< 2 times). This finding relates to the idea that there is a variety of understandings regarding the connection between research and SDGs, and this might change across SDGs (Armitage et al., 2020).

In the next analysis we will only present results using the loose threshold to allow a broader understanding of what is SDG related research. One aspect worth analysing is if the distribution of SDG related research differs across country income groups. It might be the case that some country groups are more specialised in specific SDG related research than others. In order to have a general overlook at these differences, in Fig. 2 and Fig. 3 we calculated how much SDG related research is done by researchers in countries belonging to different World Bank income groups (high, upper-middle, lower-middle and low income).

As expected, the vast majority of research is done by high and upper-middle income countries (>90%) which represent around 56% of World population during the same period. The 29 low-income countries in our set, participate in a residual share of research across SDGs, although they seem to be relatively specialised in it. Their percentage is slightly higher in SDG1 (No Poverty), SDG2 (Zero Hunger), SDG5 (Gender Equality) and SDG3 (Good health and well-being) which represents more than 50% of all the research done in the region. Regarding our 104 middle-income countries (upper and lower), they seem to be less specialised than the other regions in research related to SDGs in general, but are relatively specialised in SDG6 (Clean Water and Sanitation) and SDG7 (Affordable and Clean Energy), and less specialised in SDG5 (Gender Equality), SDG10 (Reduced Inequalities) and SDG16 (Peace, Justice and Institutions). Please note that this middle-income group includes large countries such as China, India, Brazil and Russia that during the same period comprise around 75 to 80% of World population. The other 72 high-income countries produce the majority (>50%) of research in all SDGs, but seem to be particularly specialised in SDG16 (Peace, Justice and Institutions), SDG10 (Reduced Inequalities), SDG5 (Gender Equality) and SDG4 (Quality Education).

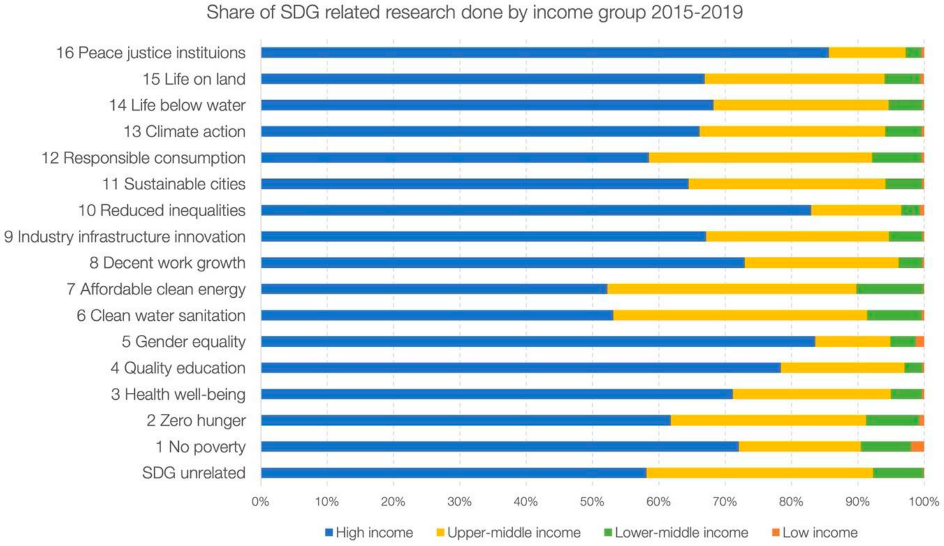


Figure 2. Share of publications associated to an SDG by income group between 2015 and 2019

Source: WoS

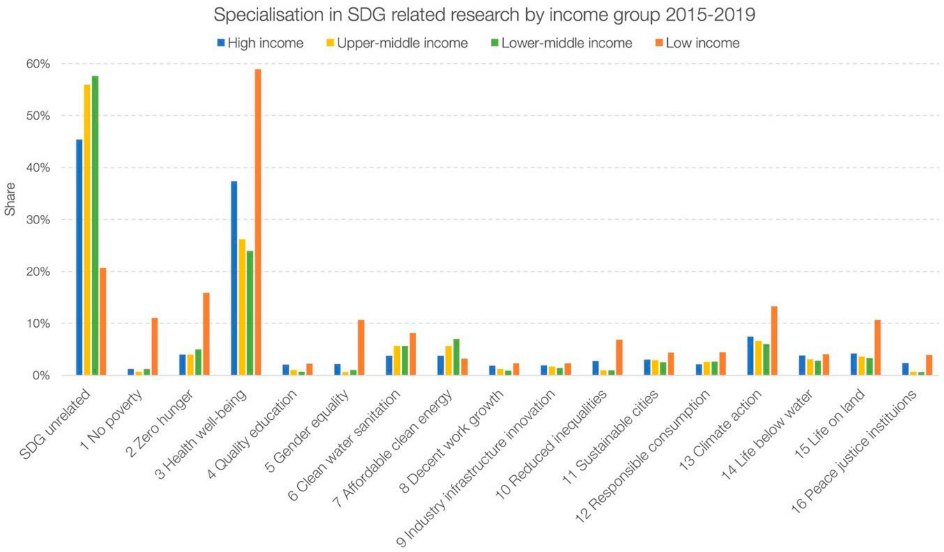


Figure 3. Revealed SDG research priorities by income group between 2015 and 2019
Source: WoS

Having analysed the research priorities in all countries across SDGs, our next step is to estimate which are the SDGs that are more problematic in each country (and country groups). Following the method illustrated in Section 2.2 for each SDG we created an index between -1 (the country is among the best performers in the world - top5%) and 1 (the country is the worst performer in the world).

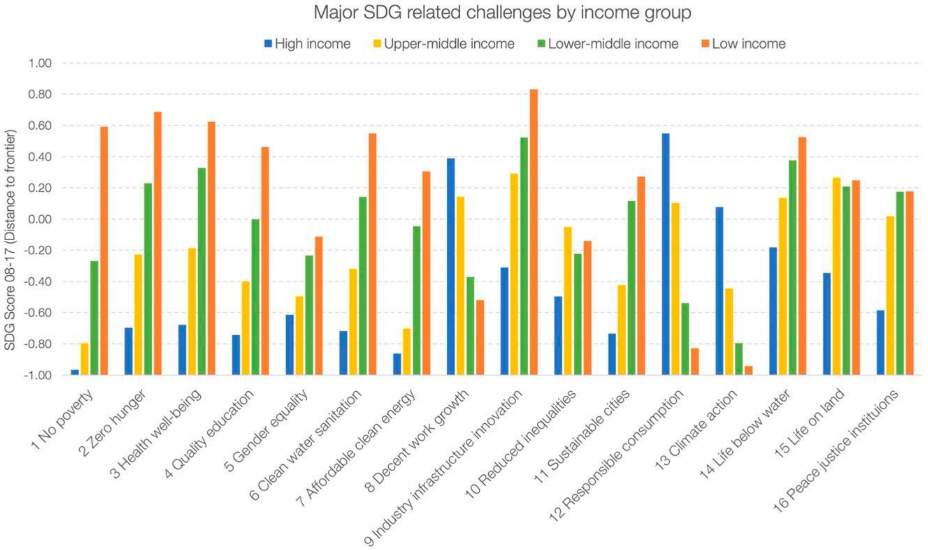


Figure 4. Revealed SDG challenges by income group between 2008 and 2017
Source: Own calculation based on UN, SDG Index and World Bank data.

Fig. 4 summarises our findings per country income group. Overall, as expected, low-income countries perform worst (further away from the frontier) in most SDGs. The exceptions are SDG12 (Responsible Consumption and Production) and SDG13 (Climate action) in which higher income countries are further away from the frontier (top 5% performers). These differences relate, for example, to the high level of domestic material consumption and waste per capita (SDG12), and high levels CO2 emissions per capita (SDG13) generated by higher income groups. Another interesting finding is that in SDGs 8 (Decent work and economic growth), 10 (Reduces inequalities) the differences between income groups are not substantial. The next question we explore is whether the SDG research specialisation of all countries, for which there is data, is related to the major challenges they face using the SDG Indicators. We use an indicator of SDG revealed research priorities discussed in section 3.1 (1 = highly specialised in research related to an SDG; -1 = no research performed related to an SDG) and an indicator of relative SDG challenge discussed in 3.2 (1 = Major challenge, country furthest away from the "frontier" in this SDG; -1 = country at the "frontier" in this SDG). Using these two dimensions we plot the scores of all countries in a specific SDG. Our major findings are the following and also derive from a pairwise correlation table in the appendix (Table A.1.):

- Linear positive correlations were found between research priorities versus countries challenges in SDG2 (Zero hunger)[7], SDG3 (Good health and well-being)[8] and SDG6 (Clean water and sanitation)[9], meaning that the countries that are further away from the frontier in these two SDGs, are the ones that are relatively more specialized in research related to these SDGs. In SDG2 and 3, these were expected findings given the long-term research specialisation that lower income countries have on Health and Agricultural Sciences (UNESCO, 2015).
- A positive but non-linear (log) correlation was also found in SDG1 (No Poverty), meaning that most countries in our sample don't suffer problems related to poverty, but the ones that do, seem to be relatively specialised in poverty related research.
- A negative correlation was found between research priorities versus countries challenges, to a certain extent in SDG4 (Quality Education), SDG7 (Affordable and Clean Energy), SDG12 (Responsible Consumption and Production), SDG13 (Climate Action). This means that the countries that are further away from the frontier in the indicators that compose these SDGs are the ones that have lower research specialisation in these SDGs.

Overall, we couldn't find a specific alignment/misalignment pattern between SDG research priorities and SDG challenges across countries. In some SDGs (1, 2, 3 and 6) there is a certain degree of alignment probably related to international research funding patterns in lower income countries (Confraria and Wang, 2020), that support research that is relevant to the major challenges in those countries. In other SDGs (e.g. 12, 13) high-income countries are, on average, further away from the frontier and they are not relatively specialised in the corresponding SDG research as the other countries.

The previous analysis gives us a broad picture, at SDG level, about levels of alignment between SDG challenges and SDG revealed research priorities in all countries. The next question we explore is whether the SDG research specialisation of specific countries is related to the major

[7] Targets related to levels of hunger, malnutrition, agricultural productivity, sustainable food production systems, genetic diversity of seeds, etc.

[8] Targets related to levels of maternal mortality, neonatal mortality, health access, disease burden, end of epidemics, etc.

[9] Targets related to wáter use efficiency, population using basic sanitation services and drinking water services, proportion of population practicing open defecation, etc.

problems they face using the SDG Indicators. Answering this question might provide guidance to rebalance research priorities towards topics closer to countries' major challenges.
As an example, we perform this analysis for three countries belonging to three different income groups (Ethiopia – Low income, India – Medium income, Germany – High income).
Our major findings in Fig. A.1. (in Appendix) are the following:

- Ethiopia shows a partial alignment. It faces significant challenges in several SDGs: SDG1 (No poverty), SDG2 (Zero hunger), SDG 3 (Good health and well-being), SDG4 (Quality education), SDG6 (Clean water and sanitation), SDG7 (Affordable and clean energy), SDG9 (Industry, infrastructure and innovation) and SDG15 (Life on land). But Ethiopia's researchers are also relatively specialised in research related to most of those SDGs. The exceptions are SDG4 (Quality education) and SDG7 (Affordable and clean energy), which are SDGs where Ethiopia has a high burden in relation to other countries.
- Our analysis shows that India faces significant challenges in SDG2 (Zero hunger), SDG 3 (Good health and well-being), SDG 5 (Gender equality), SDG6 (Clean water and sanitation), SDG11 (Sustainable cities), SDG14 (Life below water) and SDG15 (Life on land). In terms of research priorities India is relatively specialised in research related to SDG6 (Clean water and Sanitation), SDG7 (Affordable Clean Energy) and SDG12 (Responsible consumption and production). Major challenges in India such as SDG14 and SDG15 doesn't seem to receive a lot of attention in terms of relative research priorities.
- Germany seems perform worst in SDG12 (Responsible consumption and production) and SDG13 (Climate action) as most high-income countries, and their research specialisation seems to be in line with the World's average (0).

Overall, we find that higher-income countries are closer to the frontier in most SDGs (SDG 12 and 13 are exceptions) and have SDG research specialisation patterns closer to the world average. Lower-income countries present higher volatility in terms of SDG research priorities and SDG challenges.

Discussion

This paper provides a new way of exploring the extent to which countries' research priorities align to their major SDG areas of challenge. Our motivation is that there is an increasing demand for science and research funding to be better aligned with socio-economic needs, but there are very few robust methods (and datasets) that help policy makers to understand these relations. We develop an approach based on citation relations that define research areas allowing us to understand if a country is generating research (based on WoS data) relevant to their societal challenges (based on SDG indicators). We found that around two thirds of research that is done in the World in 2015-2019 is unrelated to the issues addressed in the SDGs. However, when we look at the research specialisation of specific countries by income group, we find that, on average, lower-income countries seem to perform more research on SDG issues in relative terms.
Our findings also indicate that there is a positive relation between SDG achievements by country and SDG research specialisation in specific SDGs such as SDG2 (Zero hunger) and SDG3 (Good health and well-being), meaning a certain degree of alignment between research priorities and SDG achievements across countries in these SDGs areas. However, it is essential to emphasise that these two SDGs are quite broad, and one could argue that certain types of research within those SDGs should be prioritised instead of others (Confraria and Wang, 2020; Yegros et al., 2019).

We instead found a negative relationship between research prioritisation and SDG challenges for SDG4 (Quality Education), SDG7 (Affordable and Clean Energy), SDG12 (Responsible Consumption and Production) and SDG13 (Climate Action). In the cases of SDG12 and SDG13, the countries that perform worst are high-income countries that seem to have unsustainable consumption/production patterns and contribute more to climate change. However, the countries that seem to be relatively specialised in these themes seem to be lower income countries. As for SDG 4 and 7, on average, lower-income countries seem to be the ones that face major challenges in these dimensions, being also the ones that are less specialised in research that is related to those challenges. Such misalignment between the investment in research areas and the socio-economic challenges in these SDGs/countries may reduce the possibilities of improvements in those challenges in those countries, as these countries need to rely on research capabilities developed elsewhere, where the issues faced may be different.

Lastly, this study has several limitations related to the accuracy of our estimates of SDG research priorities of countries and SDG challenges faced by them. Yet, one thing we would like to emphasise, that is particularly important for future research, is related to the marginal impact of increasing research investments in areas related to a certain SDG on the improvement of that SDG. This impact may not be the same for all SDGs. For example, local research in health (SDG3) may lead to significant improvements in the health outcomes of a country. In contrast, more local research on Poverty (SDG1) may not lead to similar marginal improvements. Future research should look carefully at this issue and consider spillovers between SDGs and positive and negative interactions between them. These may guide research prioritisation and building of research capabilities to address different challenges.

References

Armitage, C.S., Lorenz, M., Mikki, S., 2020. Mapping scholarly publications related to the Sustainable Development Goals: Do independent bibliometric approaches get the same results? *Quant. Sci. Stud.* 1, 1092–1108. doi:10.1162/qss_a_00071

Balassa, B., 1965. Trade Liberalisation and "Revealed" Comparative Advantage. *Manchester Sch*. 33, 99–123. doi:10.1111/j.1467-9957.1965.tb00050.x

Ciarli, T., Ràfols, I., 2019. The relation between research priorities and societal demands: The case of rice. *Res. Policy* 48, 949–967. doi:10.1016/j.respol.2018.10.027

Confraria, H., Wang, L., 2020. Medical research versus disease burden in Africa. *Res. Policy 49*, 103916. doi:10.1016/j.respol.2019.103916

Evans, J.A., Shim, J.-M., Ioannidis, J.P.A., 2014. Attention to Local Health Burden and the Global Disparity of Health Research. *PLoS One* 9, e90147. doi:10.1371/journal.pone.0090147

Jackson, J.E., 1991. *A Use's Guide to Principal Components*, Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA. doi:10.1002/0471725331

Rafols, I., Noyons, E., Confraria, H., Ciarli, T., 2021. *Visualising plural mappings of science for Sustainable Development Goals (SDGs)* (with Ismael Rafols, Ed Noyons and Tommaso Ciarli), in: Paper to Be Presented in the ISSI2021.

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics,* 9(1), 102–117. https://doi.org/10.1016/j.joi.2014.11.010

Sarewitz, D., Pielke Jr., R.A., 2007. The neglected heart of science policy: reconciling supply of and demand for science. *Environ. Sci. Policy*, Reconciling the Supply of and Demand for Science, with a Focus on Carbon Cycle Research 10, 5–16. doi:10.1016/j.envsci.2006.10.001

UNESCO, 2015. UNESCO Science Report: towards 2030. Paris.

Waltman, L., Van Eck, N.J., 2012. A new methodology for constructing a publication-level classification system of science. *J. Am. Soc. Inf. Sci. Technol*. 63, 2378–2392. doi:10.1002/asi.22748

Yegros, A., Van de Klippe, W., Abad-Garcia, M.F., Rafols, I., 2019. Exploring Why Global Health Needs Are Unmet by Public Research Efforts: The Potential Influences of Geography, Industry, and Publication Incentives. *SSRN Electron. J.* 2019. doi:10.2139/ssrn.3459230

Appendix.

Table A.1. Pairwise correlation of SDG challenges (2008-2017) versus SDG research priorities (2015-2019) in all SDGs

	Score_SDG1	Score_SDG2	Score_SDG3	Score_SDG4	Score_SDG5	Score_SDG6	Score_SDG7	Score_SDG8	Score_SDG9	Score_SDG10	Score_SDG11	Score_SDG12	Score_SDG13	Score_SDG14	Score_SDG15	Score_SDG16
NSI_Pubs_SDG1	0.65	0.71	0.59	0.08	0.63	0.56	-0.32	0.16	0.07	0.43	0.28	0.37	0.45	0.23	0.49	0.35
NSI_Pubs_SDG2	0.49	0.60	0.28	-0.01	0.43	0.67	-0.04	0.07	0.03	0.21	0.26	0.39	0.36	0.17	0.36	0.16
NSI_Pubs_SDG3	0.52	0.64	0.36	-0.05	0.46	0.61	-0.22	0.08	-0.04	0.26	0.16	0.31	0.34	0.16	0.40	0.23
NSI_Pubs_SDG4	0.36	0.57	0.30	-0.27	0.31	0.57	-0.12	-0.08	-0.10	0.04	0.06	0.24	0.24	0.06	0.25	-0.01
NSI_Pubs_SDG5	-0.24	-0.06	-0.09	-0.33	-0.20	0.22	0.30	-0.33	-0.17	-0.39	-0.07	-0.03	-0.32	-0.34	-0.39	-0.43
NSI_Pubs_SDG6	0.56	0.70	0.42	-0.08	0.48	0.64	-0.26	0.08	-0.03	0.28	0.20	0.35	0.44	0.25	0.50	0.21
NSI_Pubs_SDG7	0.69	0.74	0.56	0.12	0.66	0.60	-0.29	0.28	0.13	0.49	0.34	0.42	0.53	0.35	0.57	0.41
NSI_Pubs_SDG8	-0.49	-0.54	-0.27	0.10	-0.37	-0.47	0.20	-0.09	-0.01	-0.26	-0.13	-0.28	-0.30	-0.14	-0.39	-0.18
NSI_Pubs_SDG9	0.35	0.55	0.04	-0.02	0.24	0.59	-0.12	0.06	-0.05	0.09	0.17	0.22	0.39	0.23	0.37	0.08
NSI_Pubs_SDG10	0.40	0.55	0.28	0.20	0.31	0.47	-0.15	0.12	0.00	0.23	0.10	0.27	0.48	0.44	0.50	0.21
NSI_Pubs_SDG11	0.32	0.46	0.29	-0.22	0.28	0.52	-0.08	-0.07	-0.07	0.06	0.15	0.21	0.14	-0.01	0.17	-0.01
NSI_Pubs_SDG12	-0.50	-0.61	-0.34	0.13	-0.42	-0.51	0.21	-0.06	0.02	-0.22	-0.14	-0.32	-0.31	-0.14	-0.37	-0.17
NSI_Pubs_SDG13	-0.39	-0.60	-0.26	0.18	-0.30	-0.51	0.23	0.02	0.10	-0.11	-0.03	-0.26	-0.30	-0.16	-0.40	-0.07
NSI_Pubs_SDG14	0.02	0.14	-0.20	-0.13	-0.10	0.37	0.14	-0.16	-0.15	-0.20	-0.03	0.03	-0.01	-0.02	-0.02	-0.24
NSI_Pubs_SDG15	-0.10	-0.02	-0.20	-0.25	-0.07	0.14	0.12	-0.34	-0.39	-0.23	-0.19	-0.18	-0.15	-0.14	-0.16	-0.16
NSI_Pubs_SDG16	0.29	0.45	0.02	-0.19	0.07	0.48	-0.11	-0.12	-0.18	-0.07	-0.08	0.14	0.26	0.23	0.35	-0.09

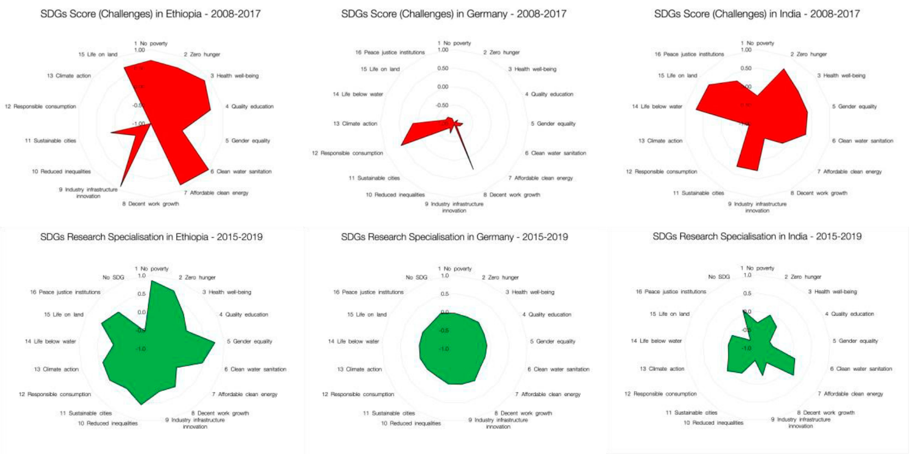


Figure A.1. Ethiopia, India and Germany main SDG challenges versus SDG research specialisation[10]

[10] Some countries don't have data available for all years in some SDG indicators. Therefore, some of the SDG Scores were not computed for some countries.

Dissension or consensus? Management and Business Research in Latin America and the Caribbean

Julián D. Cortés

julian.cortess@urosario.edu.co
Universidad del Rosario, Colombia
Fudan University, China

Abstract

This study presents longitudinal evidence on the dissension of Management and Business Research (MBR) in Latin America and the Caribbean (LAC). It looks after intellectual bridges linking clusters among such dissension. It was implemented a coword network analysis to a sample of 12,000+ articles published by authors from LAC during 1998-2017. Structural network scores showed an increasing number of keywords and mean degree but decreasing modularity and density. The intellectual bridges were those of the cluster formed by disciplines/fields that tend toward consensus (e.g., *mathematical models*) and not by core MBR subjects (e.g., *strategic planning*).

Introduction

The Winchester Mystery House has been used as an analogy for Management and Business related Research (MBR) —a purposeless although expensive entity (Davis, 2015). The hypothesis of the Hierarchy of Sciences states that while some sciences/disciplines studying simple phenomena (i.e., cells' functioning) will tend toward consensus, others studying complex phenomena (i.e., human collective behavior) will tend toward dissension (Fanelli & Glänzel, 2013). Thus, the latter path should be shaping MBR. Bottom line, dissension fragments research findings to advance on —a significant downsize effect for regions with scarce R&D resources such as Latin America and the Caribbean (LAC) (Cortés-Sánchez, 2019).

This study aims two-fold: i) to explore such dissension in MBR in LAC; ii) and to identify potential *intellectual bridges* to pooling efforts and multiply impact. I implemented the coword analysis (Callon et al., 1983). Such a science mapping (SM) technique enables stakeholders to map individual/institutional/national research and knowledge competencies (Shiffrin & Börner, 2004). Findings would be of interest to researchers in MBR, business schools, and funders by presenting the first comprehensive regional study for MBR to identify research clusters, topics as intellectual bridges between disciplines, and their evolution since 1998.

Methodology

This study modeled four coword networks for the periods 1998-2002, 2003-2007, 2008-2012, and 2013-2017. Dissension paths were explored by computing and comparing macro, meso, and micro structural network scores. Dataset and high-resolution figures are available in open access for replication and further use (Baker & Penny, 2016): http://bit.ly/2QyDJNP

Data

Bibliographic data was sourced from Scopus due to its journal coverage and LAC authors' involvement (Mongeon & Paul-Hus, 2016). The query searched for articles published from 1998 to 2017 by at least one author affiliated with any LAC institution in SCImago's *business, management, and accounting* subject area (SCImago, 2020). The final sample consisted of 12,149 articles after removing a journal with predatory features (Cortés-Sánchez, 2019) and articles with missing data. Table 1 presents the bibliometric descriptives for the top-10 LAC countries by output.

Table 1 Bibliometric descriptives for the top-10 LAC countries by output.

Total articles	12,149
Av. citation per article	14.02
Authors	20,694
Authors per document	1.7
Annual growth %	17.2

Corresponding author's countries	Articles	% Sample	MCP	Av. citation per article	Most relevant sources	% Sample
Brazil	3,256	40%	23	14.0	*Información Tecnológica*	7.8%
Mexico	874	10%	11	14.3	*J. of Cleaner Production*	5.6%
Colombia	717	8%	5	8.4	*Gestão & Produção*	4.1%
Chile	644	8%	6	17.0	*R. de Administração de Empresas*	2.4%
Argentina	427	5.3%	4	11.3	*J. of Technology Management and Innovation*	2.2%
Venezuela	129	1.6%	4	3.8	*J. of Business Research*	1.9%
Peru	96	1.2%	2	15.7	*R. Venezolana de Gerencia*	1.9%
Costa Rica	63	0.7%	0	16.4	*Estudios Gerenciales*	1.8%
Uruguay	55	0.6%	0	12.5	*R. Brasileira de Gestão de Negocios*	1.8%
Cuba	43	0.5%	0	12.9	*International J. of Production Research*	1.4%

Source: elaborated by the author based on Scopus (2020). Note: MCP: Multi-Country Publication.

Methods and software

Coword analysis enables putting together the conceptual structure based on the co-occurrence of articles keywords (Callon et al., 1983). *KeyWords Plus* method generates key-terms based on the articles' title and the references cited appearing more than twice (Clarivate Analytics, n.d.). A link connects two keywords (i.e., nodes) if both appear in the same research article (i.e., edge). The scores computed for the coword networks were: i) macro: density, mean degree, modularity; ii) meso: number of clusters; iii) and micro: betweenness (Scott & Carrington, 2014). Density is the proportion of links in a network relative to the total number of links possible. The mean degree is the average number of links per node in the network. Modularity express a networks' strength of cluster division. Increasing values indicate the existence of a community-like structure. Clustering analysis identifies highly interconnected nodes to uncover known communities. Betweenness unveils a node's capacity in mediating the flow of information in a network. Increasing values indicate a higher betweenness. Bibliometrix (Aria & Cuccurullo, 2017) and Gephi (Bastian et al., 2009) were used for networks' layout and scores' computation.

Results and discussion

Figure 1 summarizes the macro, meso, and micro scores for each network. The period 2008-2012 stands out as the network with the highest increase in the number of nodes (198%), mean degree (55%), and clusters (39%). Density and modularity, however, diminished by 50% and 2%, respectively. The networks' density and modularity had decreased throughout 1998-2017 despite the mean degree's uninterrupted growth. The number of clusters peaked in 2008-2012 (39) and reached the lowest point in 2013-2017 (22).

Ronda and Guerras (2012) found that density and clusterability, a modularity proxy, have increased between 1962-2008 for the *strategic management* coword network. That seems plausible for an MBR sub-field. Using bibliographic coupling, Fanelli and Glänzel (2013) computed a mean degree and modularity for the field of *business and economics* of 27.3 and 0.5, respectively, which were similar to the mean degree of 2008-2012 (29.6) and the modularity of 2013-2017 (0.53). The caveat is that bibliographic coupling networks are different from co-word analysis, the latter similar to that of cocitations (Yan & Ding, 2012).

Figure 2 presents the four coword networks. Nodes' size is proportional to their betweenness. Nodes with labels are: i) those among the top-10 betweenness; ii) those that increased their position among the top-20 betweenness compared to the previous period (light-green colored); and iii) those that decreased their position among the top-20 betweenness compared to the previous period (red-colored). Clusters colored in burgundy and pine-green are the first and second-largest ones, respectively. I tagged clusters manually following a discernible thread among highly connected keywords. The two most crowded clusters for each period were:

i) 1998-2002: *mathematical models, computer simulation, and algorithms* (18% nodes), and *science, technology and innovation* (STi) (10%);

ii) 2003-2007: *LAC issues* (15%), and *strategic management for sustainable development* (15%);

iii) 2008-2012: *STi in LAC* (10%), and *mathematical models, computer simulation, and algorithms* (9%); and

iv) 2013-2017: *strategic management for sustainable development* (22%); and *logistics* (14%).

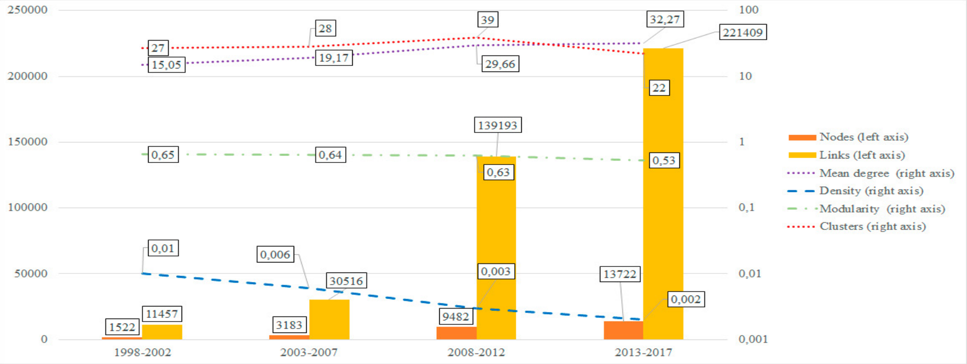


Figure 1 Nodes, links, density, modularity, and clusters of coword networks 1998-2017. Source: elaborated by the author based on Scopus (2020).

Table 2 (Appendix) presents the top-20 keywords by period according to their betweenness. Topics comprised of *mathematical models, computer simulation*, and *algorithms* cluster show consistent appearance among the four periods. *Sustainable development* and *environmental impact* related issues also showed consistent appearances. In contrast, MBR core topics, such as *strategic planning*, *quality control*, or *marketing*, showed either intermittent appearances or a downward trend.

Crowded clusters mixed between each other (e.g., STi *plus* LAC issues ⇒ STI in LAC) and specific keywords persisted (e.g., mathematical *models, computer simulation, and algorithms*; or *strategic management for sustainable development*). Contrasting those findings with the persistence and higher betweenness of the *mathematical models,* it unveils the broader applications of such methods for MBR and other disciplines (e.g., from supply chain management to e-commerce and genetics), even more than influential methods in MBR such as *case study* (actually, copied from medical sciences) (Eisenhardt, 1989). Furthermore, global solution-oriented agreements such as the Sustainable Development Goals produced inflections in the research output and permeated MBR in LAC and other developing regions (Cortés-Sánchez et al., 2020).

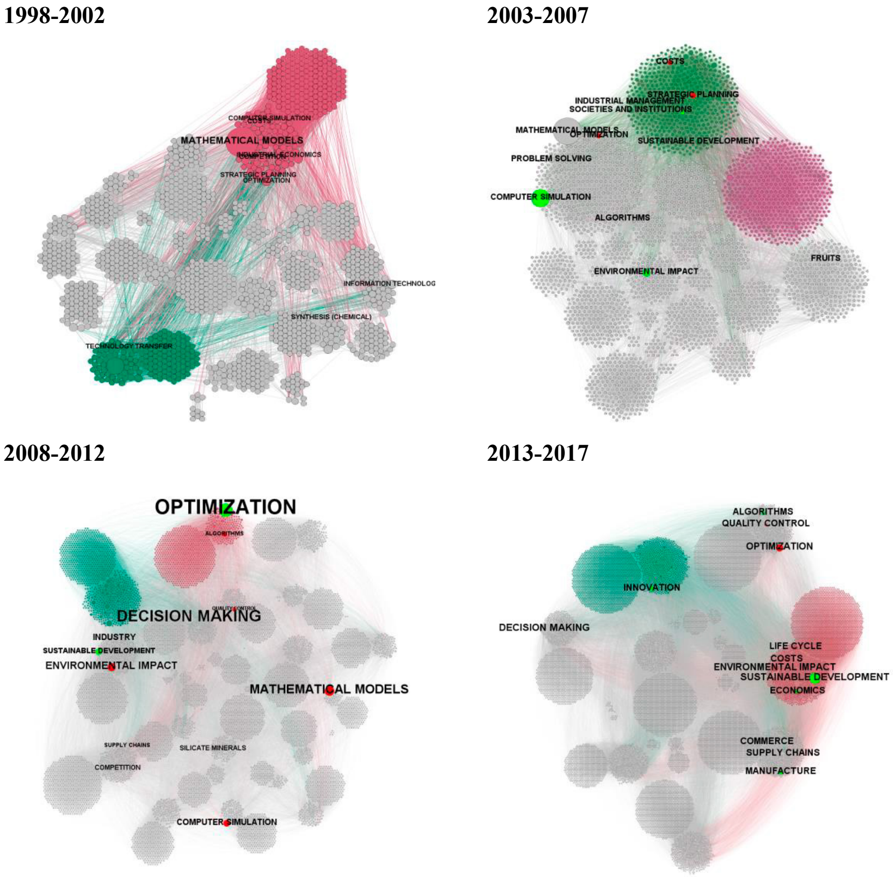


Figure 2 Coword networks 1998-2017. Layout: circle park (hierarchy 1: modularity class; hierarchy 2: betweenness). Source: elaborated by the author based on Scopus (2020) and computed with bibliometrix (2017) for R and Gephi (2009).

Conclusion

This study presented longitudinal evidence on the increasing dissension path of MBR in LAC. Also, it presented the mixture and persistence of crowded clusters and intellectual bridges. Such bridges were not from MBR but different disciplines moving toward consensus (e.g., mathematics or physics). Further research could implement non-redundant SM techniques such as social (i.e., coauthorships) or information (i.e., bibliographic coupling) based networks, discuss the advantage and obstacles of open vs. closed network structures, and source data from broader and inclusive bibliographic databases (e.g., Google Scholar or Dimensions).

References

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Baker, M., & Penny, D. (2016). Is there a reproducibility crisis? *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452A

Bastian M., Heymann S., & Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*. https://gephi.org/users/publications/

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Baker, M., & Penny, D. (2016). Is there a reproducibility crisis? *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452A

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*. https://gephi.org/users/publications/

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, *22*(2), 191–235. https://doi.org/10.1177/053901883022002003

Clarivate Analytics. (n.d.). *KeyWords Plus generation, creation, and changes*. Retrieved October 8, 2020, from https://bit.ly/3ouh9m4

Cortés-Sánchez, J. D. (2019). Innovation in Latin America through the lens of bibliometrics: crammed and fading away. *Scientometrics*, *121*(2), 869–895. https://doi.org/10.1007/s11192-019-03201-0

Cortés-Sánchez, J. D., Bohle, K., & Guix, M. (2020). *Innovation Research in Management and STEM for Sustainability in Developing Countries Insights from Bibliometrics in the Global South* (No. 154; Universidad Del Rosario, School of Management). https://repository.urosario.edu.co/handle/10336/28225

Davis. (2015). What Is Organizational Research For? *Administrative Science Quarterly*, *60*(2), 179–188. https://doi.org/10.1177/0001839215585725

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *The Academy of Management Review*, *14*(4), 532–550. https://doi.org/10.2307/258557

Fanelli, D., & Glänzel, W. (2013). Bibliometric Evidence for a Hierarchy of the Sciences. *PLOS ONE*, *8*(6), e66938. https://doi.org/10.1371/journal.pone.0066938

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Ronda-Pupo, G. A., & Guerras-Martin, L. A. (2012). Dynamics of the evolution of the strategy concept 1962-2008: A co-word analysis. *Strategic Management Journal*, *33*(2), 162–188. https://doi.org/10.1002/smj.948

SCImago. (2020). *SJR : Scientific Journal Rankings*. https://www.scimagojr.com/journalrank.php

Scopus. (2020). *Scopus - Document search*. https://www.scopus.com/search/form.uri?display=basic

Scott, J., & Carrington, P. (Eds.). (2014). *The SAGE Handbook of Social Network Analysis*. SAGE Publications Ltd. https://doi.org/https://dx.doi.org/10.4135/9781446294413

Shiffrin, R. M., & Börner, K. (2004). Mapping knowledge domains. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(SUPPL. 1), 5183–5185. https://doi.org/10.1073/pnas.0307852100

Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326. https://doi.org/10.1002/asi.22680

Appendix

Table 2 Top-20 keywords based on betweenness score 1998-2017

1998-2002		2003-2007		2008-2012		2013-2017	
Keyword	Betweenness	Keyword	Betweenness	Keyword	Betweenness	Keyword	Betweenness
mathematical models	2,62e+11	mathematical models =	1,10e+12	optimization ↑	2,62e+12	sustainable development ↑	6,52e+12
costs	5,61e+10	computer simulation ↑	6,95e+11	***decision making***	2,15e+12	decision making =	6,02e+12
computer simulation	5,27e+10	environmental impact ↑	2,28e+11	mathematical models ↓	1,83e+12	optimization ↓	4,12e+12
strategic planning	5,14e+10	***industrial management***	1,62e+11	environmental impact ↓	1,48e+12	environmental impact =	3,65e+12
argentina	5,03e+10	***problem solving***	1,59e+11	***industry***	1,25e+12	manufacture ↑	2,69e+12
synthesis (chemical)	4,35e+10	***north america***	1,56e+11	computer simulation ↓	1,20e+12	***life cycle***	2,60e+12
optimization	4,07e+10	***algorithms***	1,55e+11	sustainable development ↑	1,10e+12	innovation ↑	2,37e+12
technology transfer	3,81e+10	strategic planning ↓	1,40e+11	***colombia***	9,88e+11	***costs***	2,35e+12
information technology	3,59e+10	***sustainable development***	1,40e+11	***competition***	9,75e+11	***supply chains***	2,33e+12
industrial economics	3,29e+10	***quality control***	1,35e+11	***silicate minerals***	9,54e+11	***commerce***	2,28e+12
environmental impact	3,25e+10	***data acquisition***	1,33e+11	***modeling***	8,45e+11	algorithms ↑	2,21e+12
water	3,19e+10	***fruits***	1,31e+11	algorithms ↓	8,43e+11	economics ↑	2,20e+12
structural analysis	3,17e+10	marketing ↑	1,30e+11	***production engineering***	8,06e+11	***environmental management***	1,57e+12
thermal effects	3,06e+10	***societies and institutions***	1,24e+11	***concentration (process)***	7,29e+11	***carbon dioxide***	1,50e+12
performance	2,96e+10	optimization ↓	1,13e+11	***economics***	7,20e+11	***integer programming***	1,32e+12
marketing	2,71e+10	***food processing***	1,09e+11	***manufacture***	7,03e+11	***education***	1,28e+12
management	2,58e+10	costs ↓	1,06e+11	quality control ↓	6,96e+11	***regression analysis***	1,25e+12
raw materials	2,56e+10	***scheduling***	9,66e+10	***innovation***	6,82e+11	***investments***	1,23e+12
internet	2,52e+10	***systems analysis***	9,13e+10	***simulation***	6,64e+11	quality control ↓	1,21e+12
venezuela	2,34e+10	***public policy***	9,00e+10	***developing countries***	6,39e+11	***fruits***	1,19e+12

Note: New keywords compared to the former period are **bold** and *italic*. Symbols tell if the keyword increased (↑), diminished (↓), or maintained (=) its rank compared to the previous period. Source: elaborated by the author based on Scopus (2020) and computed with bibliometrix (2017) for R and Gephi (2009)

ISBNs as identifiers for books in research evaluation

Eleonora Dagienė[1] and Kai Li[2]

[1] *e.dagiene@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University,
PO Box 905, 2300 AX Leiden, The Netherlands
Faculty of Human and Social Studies, Mykolas Romeris University,
Ateities g. 20, LT-08303 Vilnius, Lithuania

[2] *kai.li@ruc.edu.cn*
School of Information Resource Management, Renmin University of China,
59 Zhongguancun St, Haidian District, Beijing, China, 100872

Abstract

Books and evaluations thereof are essential components of the research system. Despite existing efforts to use citation-based metrics to evaluate books, we argue that a more effective book evaluation system should focus on the entire lifecycle of book publication, not just the final products. Building on this idea, we here present the preliminary results of an ongoing project aimed at establishing such a system. As the first step of this project, we analysed ISBNs used for book outputs submitted to two national research evaluation systems, the UK Research Excellence Framework 2014 and the formal Lithuanian research assessment for 2004–2016. By conducting this analysis, we investigate whether and how ISBNs might be used as a reliable identifier in collecting bibliographic data, which will then serve as the foundation of the new book evaluation system. We describe our approach to ISBN collection, processing and analysis, and discuss the suitability of ISBNs for our purposes. Notably, we found ISBNs a handy indicator, given their relevant ubiquity in the book-centric bibliographic universe. However, processing ISBNs from existing systems is very time-consuming and requires extensive manual effort, which calls for a better book bibliographic data infrastructure.

Introduction

The evaluation of books is critical to the research assessment system because books are an important type of research output (Giménez-Toledo, Mañana-Rodríguez, & Tejada-Artigas, 2015; Kulczycki et al., 2018; Zuccala & Robinson-García, 2019). Usually, book evaluation is carried out by experts, who either assess book publishers or review individual books (Giménez-Toledo et al., 2019). However, ambiguous rules in book publishers' rankings (Dagienė, 2020) and institutional competition for funding have led to gaming the system (Aagaard, 2015; Mouritzen & Opstrup, 2020; Rowlands & Wright, 2019). Several studies have suggested that book metrics can help counteract a certain degree of academic bias in book assessment (Faggiolani & Solimine, 2018; Williams, Basso, Galleron, & Lippiello, 2018).

Quantitative indicators for books have been gaining momentum in recent years. Various approaches have been taken to investigate these metrics; most focus on the ex-post assessment of books and their impact (Liu, Ding, & Gu, 2017; Torres-Salinas, Robinson-Garcia, & Gorraiz, 2017; White & Zuccala, 2018; Zhou & Zhang, 2020). Thus far, it seems that only evaluation labels proposed for identifying peer-reviewed publications (Kulczycki et al., 2019) provide an ex-ante solution for measuring book quality.

After investigating criteria, processes, and practices for assessing book publishers across countries, Dagienė (2020) suggested a new approach to book evaluation. This model accounts for the value that book producers add to a manuscript at different publishing stages (e.g., quality control, production, dissemination, archiving, and marketing)—assuming here that it is the stages that are deemed essential for effective scholarly communication (Publications Office of the European Union, 2019).

Different publishers do not contribute equally to each step of book production. For example, (1) not all publishers conduct peer review, so there is a need for unique labels (Kulczycki et al., 2019); (2) many publishers do not edit submitted manuscripts, which they deem the authors' responsibility; (3) some publishers do not consider long-term archiving necessary, and some publishers do not consider digital formats necessary at all. Still other essential parts of book production important for the future of scholarly communication (Kraker et al., 2016), such as open access or understandability, are not even considered in research evaluation.

Ideally, publishers need to offer information on their contribution in each step as part of the book metadata. The availability of such information would allow book outputs to be evaluated without relying on general book publisher information (such as editorial board members, peer-review procedures, and the publishing programme).

Moreover, since this new approach to book assessment works on the book (not publisher) level, we suppose that the ISBN Standard (already established in the book industry) will be the core of this approach. However, such a system's optimal implementation process still needs to be considered from many perspectives.

This research-in-progress paper addresses only one goal of the ongoing project, namely, assessment of the use of ISBNs as the primary book identifier for such a system. The broader aims of our project are: (1) to determine how much metadata about books is available in various sources; (2) to identify the publishers of books submitted by research institutions as their research outputs; and (3) to examine the extent to which these publishers have provided high-quality metadata at the level of individual publications.

In this paper, we pursued the following questions as initial steps in the above research agenda:
(1) How many books submitted to research evaluations have ISBNs that can be used to gather their metadata?
(2) What are the reasons for missing ISBNs?
(3) How can the use of ISBNs in research evaluations be improved?

This paper provides the results of our efforts to obtain book metadata based on the ISBNs collected in national research evaluation exercises. We also report issues in collecting and processing ISBN data.

Having a collection of valid ISBNs is an essential step towards constructing a book evaluation system that focuses on the individual components of the publication workflow and can be implemented automatically.

Research design

Many researchers have examined the sources of books' bibliographic data (Halevi, Nicolas, & Bar-Ilan, 2016; Torres-Salinas & Moed, 2009) or book metrics (White et al., 2009; Zhu, Yan, Peroni, & Che, 2020). Unfortunately, in the databases explored to date, it is impossible to obtain a comprehensive sample of books and publishers, so as to evaluate all possible variations of book outputs and their publishers at the country level. In this study, we used openly available data on books already submitted by academic institutions as part of national research evaluation exercises. Specifically, we used data from two evaluation systems that recognise books as institutional research outputs and empower experts to assess submitted books at the individual level. These are the UK Research Excellence Framework 2014 (REF 2014)[1] and the formal Lithuanian annual assessment, whose data span two databases ranging from 2004 to 2016 (Dagienė, 2020; Dagienė, Kriščiūnas, Tautkevičienė, & Maskeliūnas, 2019).

Bibliographic information on Lithuanian book outputs was derived from databases managed by the Lithuanian Research Council: (1) Dynamics of Lithuanian Research Potential[2] for outputs published 2004–2008 and (2) Reports on Scientific, Arts and other Relevant Activities of Research and Higher Education Institutions[3] for outputs published 2009–2016.

[1] Research Excellence Framework 2014. https://results.ref.ac.uk accessed 22 December 2020.
[2] Lietuvos mokslo potencialo dinamika http://www.mokslas.mii.lt/mokslas/ accessed 16 April 2020
[3] Mokslo ir studijų institucijų mokslinės, meninės ir su jomis susijusios kitos veiklos ataskaita https://mokslas.lmt.lt/INSTITUCIJOS/ accessed 16 April 2020

From the REF website, we downloaded all book ISBNs, which were further validated and parsed using the *isbnlib*[4] Python package. We also used the plugin *isbnlib-worldcat* [5] to discover the numbers of books included in WorldCat and obtain their initial metadata.

The prefixes extracted from the ISBN codes identify the publishers. The *Global Register of Publishers (GRP)*[6], created by the International ISBN Agency, contains the country and the agency responsible for the ISBNs' registration as well as detailed information about the publisher: the exact name, current status (active/inactive), address, website, and administration's contact information (name, phone, fax, and email). However, in this study, we used only a portion of this available data to analyse patterns among publishers and their imprints/brands.

Searching for valid ISBNs and other attributes of book outputs

REF 2014 policies describe ISBNs as optional identifiers. Hence, institutions submitting bibliographic data for their book outputs choose to submit an ISBN (and, as various correct ways exist, choose the method for presenting the ISBN), or they skip this field when submitting the entry. However, in Lithuania, the formal research assessment regulations mandate ISBNs for any book output. Nevertheless, both British and Lithuanian records contained sufficient ISBNs to investigate the amount and quality of metadata available at the book level. Before using these acquired ISBNs to collect metadata, we first checked their validity. Table 1 shows the numbers of records collected as initial data (Stage 1) and those remaining after data cleaning (Stage 2). The methods and results of this process are presented next.

Cleaning and unifying ISBNs in the UK REF 2014 book output data

We validated all gathered ISBNs using the *isbnlib* Python package and manually searched for missing or incorrectly presented ISBNs in various online catalogues (Stage 2). Table 1 summarises the results of extracting unique and valid ISBNs from REF 2014 data.

Table 1. Numbers of records with ISBNs before and after cleaning (REF 2014)

Type of outputs in REF 2014	*Stage in data cleaning*	*Records in REF 2014*	*Unique entries*				*Linked only with ISSNs*	*Not applicable for ISBNs*	*Other reasons*
			total	*valid*	*not valid*	*missing*			
(A) Books	Stage 1	10,371	10,173	10,161	12	8		10	
	Stage 2		10,064	10,064	–	–	7	2	6
(B) Edited books	Stage 1	2,133	2,022	2,011	11	11		37	
	Stage 2		1,962	1,962	–	–	36	1	1
(C) Chapters in edited volumes	Stage 1	14,396	11,771	11,748	24	35	9	18	
	Stage 2		10,401	10,401	_	–	8	2	14

Since an ISBN was an optional, manually added data element, those ISBNs that were submitted at all varied greatly in presentation. Ten and thirteen-digit ISBNs of the same book (isbn10 or isbn13) were sometimes provided, dashes were included in some records and omitted in others, typographical errors appeared, and ISSNs were sometimes used instead of ISBNs. When we cleaned the records by removing unnecessary characters or spaces and converting isbn10 codes to isbn13, the list of unique ISBNs shortened.

[4] Python library isbnlib 3.10.4. https://pypi.org/project/isbnlib/ accessed 12 January 2021.

[5] WorldCat is the world's largest library catalogue. https://www.worldcat.org/ accessed 12 January 2021.

[6] The Global Register of Publishers https://grp.isbn-international.org/ accessed 22 December 2020.

The figures in the initial REF 2014 data sets differ from the counts of valid ISBNs because different institutions submitted the same co-authored books. A major source of these differences is the edited volume resources. For example, some ISBNs (attributed to edited volumes) combined eight entries (chapters) of the same book in the initial dataset. After data cleaning and ISBN validation, it appeared that not eight, but 15 entries (chapters) belonged to the same ISBN (i.e., the same edited volume). In the end, the number of edited volumes shrank from 11,771 to 10,401.

Reasons for missing ISBNs in REF 2014

Our data cleaning process left 77 entries without ISBNs. We found three main reasons for this omission:

Series and journals with ISSNs. Most outputs in this category were submitted under *edited books* (Table 1, type B). These can be books or volumes in continuously published book series; alternatively, they can be 'special issues' in journals with their own ISSNs. For these resources, it is technically permissible for individual books to be assigned ISBNs in addition to ISSNs (*ISBN users' manual*, 2017). In analysing the data and searching for missing ISBNs, we have noticed that some academic publishers use this practice.

ISBNs are not applicable. This category consists of publications that are not eligible for ISBNs under the ISBN Standard. These outputs include exhibition catalogues (printed or issued online), dictionaries (continuously published online), teaching content (issued for educational purposes and not for sale), and personalised books not intended for general availability or purchase.

Other reasons. We used this last category to include outputs which could have had ISBNs but did not. While the exact reason for this omission was unknown for each case, we identified several different scenarios. First, some PDFs were made publicly available on institutional repositories as books without ISBNs. Authors could self-publish—an option recognised in the ISBN Users' Manual (2017, 3.16)—and thus order ISBNs for their own books. Second, several publishers in the REF 2014 sample were inconsistent in assigning ISBNs, providing codes for some books in their catalogues but not for others. We cannot find any explanation for this in the publishers' book catalogues or websites. Third, and most unusually, publications produced by Helle Panke lack ISBNs 'as the organisation does not wish to be bound to Germany's regulations concerning book pricing.'[7] The fourth scenario is that publishers printed poorly formed ISBNs on the books' copyright pages, and no other valid ISBNs were found for these books. Lastly, some invalid ISBNs appeared only in the REF 2014 records and institutional repositories; we could not find bibliographic information about either these books or their ISBN codes in any library or union catalogue.

The UK final dataset contains 22,427 valid ISBNs representing 26,823 book-related outputs submitted to REF 2014.

ISBNs in Lithuanian data on book outputs

The first dataset gathered from the Lithuanian Research Council databases included bibliographic information on 4,430 book outputs submitted from 2004 to 2016. Similar issues and scenarios arose during the cleaning of the Lithuanian data as had appeared with REF 2014 (Dagienė, 2020; Dagienė et al., 2019), and similar challenges were encountered regarding the validity of ISBN codes. The *isbnlib* package revealed 65 invalid ISBNs in the initial data set; we consulted the Lithuanian Academic Electronic Library[8] holdings and found valid codes for 44 (out of 65) missing ISBNs. However, we also found 21 invalid ISBNs included in some

[7] Bibliographical note in University of Bristol repository. https://research-information.bris.ac.uk/en/publications/the-party-education-year-of-the-socialist-unity-party-of-germany- accessed 12 January 2021.

[8] The Lithuanian Academic Electronic Library https://www.lvb.lt/ accessed 12 January 2021.

books' bibliographic records. After all, because we did not find any other ISBNs for 19 books, we excluded them from our final dataset of 4,074 records.

Discussion and conclusions

In this paper, we present the preliminary results of an ongoing project aimed at establishing a new approach to book evaluation. We found processing ISBNs from existing systems to be very time-consuming and require significant manual effort.

Examining submissions to two national research evaluation exercises (in the UK and Lithuania), we found that nearly all book outputs have ISBNs. This suggests that an ISBN is a suitable identifier for gathering the metadata needed to construct and demonstrate the approach proposed in our project.

Although almost all book outputs have ISBNs, our results show that many ISBNs in book evaluation systems or library catalogues have not been validated, which makes it more difficult for the data to be used or reused automatically. We found more than 1,500 entries with incorrect or missing ISBNs, which required tremendous data cleaning efforts. To address this matter, we recommend that book evaluation systems implement an automatic ISBN validation system at the submission stage, like that undertaken in the REF 2021 process[9]. Such a mechanism would benefit other book information systems, such as library catalogues and book output registration systems (e.g., institutional repositories or systems for administration of research assessment exercises). This change will improve our bibliographic universe by making higher-quality data available right at the outset of data collection. The work reported in this paper represents the first step towards this goal.

As the next step of our project, we are planning to use the collected ISBNs to gather metadata records for books in the book evaluation systems. These records will be used to construct a new book evaluation ecosystem focusing on how book publishers participate in the publication lifecycle, which will produce more meaningful results about the quality of books and their publishers.

References

Aagaard, K. (2015). How incentives trickle down: Local use of a national bibliometric indicator system. *Science and Public Policy*, *42*(5), 725–737. https://doi.org/10.1093/scipol/scu087

Dagienė, E. (2020). *Prestige of scholarly book publishers – an investigation into criteria, processes, and practices across countries* (Vol. XX). Retrieved from https://arxiv.org/abs/2008.06008

Dagienė, E., Kriščiūnas, A., Tautkevičienė, G., & Maskeliūnas, S. (2019). Impact of national research assessment exercises on monographs and scholarly books authored by the Lithuanian researchers. In *ISSI 2019: 17th International Conference of the International Society for Scientometrics and Informetrics, 2-5 September, Rome, Italy* (pp. 2716–2717). Leuven: International Society for Scientometrics and Informetrics. Retrieved from http://www.issi-society.org/publications/issi-conference-proceedings/proceedings-of-issi-2019/

Faggiolani, C., & Solimine, G. (2018). Mapping the role of the book in evaluation at the individual and department level in Italian SSH. A multisource analysis. In *The Evaluation of Research in Social Sciences and Humanities* (pp. 33–53). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-68554-0_2

Giménez-Toledo, E., Mañana-Rodríguez, J., Engels, T. C. E., Guns, R., Kulczycki, E., Ochsner, M., … Zuccala, A. A. (2019). Taking scholarly books into account, part II: a comparison of 19 European countries in evaluation and funding. *Scientometrics*, *118*(1), 233–251. https://doi.org/10.1007/s11192-018-2956-7

[9] REF 2021 submission system validation rules https://www.ref.ac.uk/media/1665/submissions-system-validation-documentation-for-ref2021-dec2020.pdf accessed 22 December 2020

Giménez-Toledo, E., Mañana-Rodríguez, J., & Tejada-Artigas, C.-M. (2015). Review of national and international initiatives on books and book publishers assessment. *El Profesional de La Información*, *24*(6), 705. https://doi.org/10.3145/epi.2015.nov.02

Halevi, G., Nicolas, B., & Bar-Ilan, J. (2016). The complexity of measuring the impact of books. *Publishing Research Quarterly*, *32*(3), 187–200. https://doi.org/10.1007/s12109-016-9464-5

ISBN users' manual. (2017) (Seventh Ed). London: International ISBN Agency.

Kraker, P., Dörler, D., Ferus, A., Gutounig, R., Heigl, F., Kaier, C., … Vignoli, M. The Vienna Principles: A Vision for Scholarly Communication in the 21st Century (2016). https://doi.org/http://doi.org/10.5281/zenodo.55597

Kulczycki, E., Engels, T. C. E. E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., … Zuccala, A. (2018). Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics*, *116*(1), 463–486. https://doi.org/10.1007/s11192-018-2711-0

Kulczycki, E., Rozkosz, E. A., Engels, T. C. E. E., Guns, R., Hołowiecki, M., & Pölönen, J. (2019). How to identify peer-reviewed publications: Open-identity labels in scholarly book publishing. *PLOS ONE*, *14*(3), e0214423. https://doi.org/10.1371/journal.pone.0214423

Liu, W., Ding, Y., & Gu, M. (2017). Book reviews in academic journals: patterns and dynamics. *Scientometrics*, *110*(1), 355–364. https://doi.org/10.1007/s11192-016-2172-2

Mouritzen, P. E., & Opstrup, N. (2020). *Performance Management at Universities*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-21325-1

Publications Office of the European Union. (2019). *Future of scholarly publishing and scholarly communication. European Commission*. https://doi.org/10.2777/836532

Rowlands, J., & Wright, S. (2019). Hunting for points: the effects of research assessment on research practice. *Studies in Higher Education*, *0*(0), 1–15. https://doi.org/10.1080/03075079.2019.1706077

Torres-Salinas, D., & Moed, H. F. (2009). Library catalog analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in economics. *Journal of Informetrics*, *3*(1), 9–26. https://doi.org/10.1016/j.joi.2008.10.002

Torres-Salinas, D., Robinson-Garcia, N., & Gorraiz, J. (2017). Filling the citation gap: measuring the multidimensional impact of the academic book at institutional level with PlumX. *Scientometrics*, *113*(3), 1371–1384. https://doi.org/10.1007/s11192-017-2539-z

White, H. D., Boell, S. K., Yu, H., Davis, M., Wilson, C. S., & Cole, F. T. H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology*, *60*(6), 1083–1096. https://doi.org/10.1002/asi.21045

White, H. D., & Zuccala, A. A. (2018). Libcitations, WorldCat, cultural impact, and fame. *Journal of the Association for Information Science and Technology*, *69*(12), 1502–1512. https://doi.org/10.1002/asi.24064

Williams, G., Basso, A., Galleron, I., & Lippiello, T. (2018). More, less or better: The problem of evaluating books in SSH research. In *The evaluation of research in social sciences and humanities* (pp. 133–158). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-68554-0_6

Zhou, Q., & Zhang, C. (2020). Evaluating wider impacts of books via fine-grained mining on citation literatures. *Scientometrics*, *125*(3), 1923–1948. https://doi.org/10.1007/s11192-020-03676-2

Zhu, Y., Yan, E., Peroni, S., & Che, C. (2020). Nine million book items and eleven million citations: a study of book-based scholarly communication using OpenCitations. *Scientometrics*, *122*(2), 1097–1112. https://doi.org/10.1007/s11192-019-03311-9

Zuccala, A., & Robinson-García, N. (2019). Reviewing, Indicating, and Counting Books for Modern Research Evaluation Systems. In *Springer Handbook of Science and Technology Indicators* (pp. 715–728). https://doi.org/10.1007/978-3-030-02511-3_27

An Ontology-Based Semantic Design for Good Evaluations of Research Practices

Cinzia Daraio[1] and Alessio Vaccari[2]

[1] *daraio@diag.uniroma1.it*
Department of Computer, Control and Management Engineering "Antonio Ruberti" (DIAG), Sapienza University of Rome, Rome (Italy*)*

[2] *alessio.vaccari@uniroma1.it*
Department of Philosophy, Sapienza University of Rome, Rome (Italy*)*

Abstract

The main objective of this paper is to develop a knowledge base for the development of "good" evaluations of research practices. A good evaluation of research practices, intended as social practices à la MacIntyre, should take into account the stable motivations and the traits of the characters (i.e. the *virtues*) of researchers. The development of a questionnaire for the assessment of researchers' virtues is a challenging task. The use of ontologies and taxonomic knowledge, and the reasoning algorithms that can make inferences on the basis of such knowledge represents a way for testing the consistency of the information reported in the questionnaire and to analyze in a correct and coherent way the data gathered through it.

Introduction

We live in an evaluation society (Dahler-Larsen, 2011): evaluations, quantifications and assessments are everywhere and analyse, scrutinize and monitor all kind of aspects of scholarly activities. In addition, in recent decades, the rapid changes taking place in the production, communication and evaluation of research have been signs of an ongoing transformation. Largely, we are facing a transition from a traditional evaluation model, based on indicators (e.g. number of publications and citations) to a modern evaluation, characterized by a multiplicity of distinct, complementary and new dimensions including the so called *altmetrics*. This situation is led by the development and increasing availability of data and statistical and computerized techniques for their treatment, including among others the recent advancements in artificial intelligence and machine learning.

In this context, it becomes increasingly important to "evaluate evaluations", beyond the convenience calculations of the past (Larson and Berliner, 1983), looking for "good" evaluations. Although there is a proliferation of increasingly sophisticated quantitative methods to evaluate research, there is still a lack of clarity on how to understand and operationalize the notion of "good" evaluation of research practices. Researchers' motivations and their specific traits of character (i.e. researchers' *virtues*) have recently been recognized as important factors that must be considered to make a "good" evaluation (Daraio and Vaccari, 2020). Good evaluation takes into account the constitutive elements of *research practices*, intended as *social practices*, according to the MacIntyre view (MacIntyre, 1985; Murdoch, 1998). Following this line, good evaluation must consider researcher's virtues in the realization of the "internal goods" of the social practice they are involved in.

The main result of this recent study (Daraio and Vaccari, 2020) is the preliminary version of a questionnaire to assess the researcher's virtues, to be able to complement current research evaluations with additional elements referring to scholar's motivations and personal threats of character currently excluded by bibliometric indicators in use.

The main objective of this paper is to exploit the properties of ontology-based modelling to represent the domain of researchers' virtues in order to prepare a questionnaire to be administered to researchers. This approach will allow us 1) to check the consistency and the coherence of the questionnaire content and structure before its application (or use) and 2) to correctly and coherently interpret the data gathered through the questionnaire.

The paper unfolds as follows. The next section illustrates the main problems that have to be addressed for the assessment of researchers' virtues. The following section introduces the virtues of researchers as starting point for the semantic modelling of the domain and further elaborates on the questionnaire proposed in Daraio and Vaccari (2020). Section 4 outlines the ontology-based modelling. Section 5 summarizes existing literature on the topics and the last section concludes the paper.

Towards the assessment of researchers' virtues: challenges and first steps

The assessment of researchers' virtues is challenged by several pitfalls and problems. The first issue is related to the question of the measurement of virtues. Virtue, it has been argued, seems to consist of a special sensitivity that escapes empirical measurement (Murdoch, 1998). Recently, however, some scholars have tried to undermine this pessimist assumption by giving hope to those of us who seek to develop a model of evaluation of research including also a component based on the virtues. Snow (2014) has recently launched a promising line of research based on the elements of psychology that characterize the virtues. In its perspective *virtue* is composed of the following three elements:

1) *intelligence*, which highlights the fact that virtue proceeds from a set of cognitive and emotional mental states that enable us to be sensitive to some morally relevant features of the situations in which, really or imaginatively, we find ourselves (Snow, 2014, pp. 4-5). See also Snow (2010 and 2012);
2) *dispositionality*, refers to the fact that this state is a trait of the personality of the agent and is not an occasional element of his psychology;
3) *behaviour*, i.e. virtue typically manifests itself in the actions and other behavioural responses of the virtuous person (Snow, 2010, pp. 4-5).

Snow argues that each of these characteristics of virtue can be measured and she outlines a model that consists of three measurement criteria. First, the agent's performance must be taken into account, i.e. the presence of the virtue in question must be verified from the agent's ability to repeatedly perform a given behavioural pattern in the different situations that constitute, so to speak, the field of action of a specific virtue. Secondly, Snow believes it is crucial to take into account the reports that agents make of their emotional and cognitive life during the performance of actions that they consider virtuous. To facilitate this task, Snow believes it is desirable that, on the model of some US colleges, research institutions make available to their participants' special apps that can be downloaded on any electronic device, allowing them to collect the results of the self-observations of agents. Gathering the products of introspection, in addition to offering a useful material to those who are called to assess the presence of virtues in others, also allow agents to take into account the health of their virtues and measure any flexing or, on the contrary, increases in their readiness and effectiveness in responding to the pressures the world exerts on them. Finally, Snow argues that it is important to connect these data with those that impartial observers, in the figure of external evaluators, can collect in the course of annual surveys covering both the outputs of the research and the way in which the researcher dwells in different spheres of social interaction with other participants in the practice.

A further problem to be addressed is which questions to introduce in the questionnaires. These must be sufficiently diversified to allow the evaluators to answer not only the dry question about whether or not there is a virtue, but also to determine the *quantum* of it. Snow suggested four levels to be introduced in the questionnaires.

I. The first verifies the presence in the agent of receptivity to the *stimulus* that typically activates virtue.

II. The second examines its ability to recognize the virtue appropriate to the given circumstance.

III. The third verifies the most complex ability to generate a virtuous response.

IV. The fourth, finally, measures the ability of the agent to generate the virtuous response in a cross-cutting way to a plurality of situations.

Following the four levels questions introduced by Snow, it is possible to measure on a scale from 0 (minimum) to 4 (maximum) the researcher's mastery of virtue. This is done over a spectrum ranging from *(1)* the ability to understand the importance of the problem to which virtue constitutes an answer, to *(2)* the ability to recognize the virtue in question, to *(3)* the ability to express virtue occasionally, to *(4)* the ability to manifest it in all situations that constitute the scope of that virtue.

On top of these problems, the traditional problems related to the development of questionnaire and the collection of the necessary information through questionnaire and interview arise (Hochschild, 2009; Rabionet, 2011; Kvale, 2008 and Wolcott, 2008).

In this paper we suggest to exploit the advantages of an ontology-based modelling that we will illustrate after the next section, for overcoming the aforementioned problems.

3. Semantic modelling of the virtues of researchers

The starting point for the semantic modelling of the domain in exam are the virtues of researchers.

Let's start with a general characterization of the nature of virtue. Virtues are dispositions to believe, to feel emotions and act in certain ways that are activated when we perceive relevant characteristics in the world. A courageous person, for example, will face danger firmly when she believes that someone or something to which she attaches a value is at risk of being harmed. Similarly, a charitable person when she believes that someone is suffering will tend to sympathize with that suffering and, consequently, take care of it. Furthermore, an accurate person is one who does not accept any belief-shaped thing that enters his or her mind, but takes special care in forming his or her beliefs. Actions of these type will not be singular but will be stable and predictable: whenever she perceives certain relevant characteristics of a situation, the virtuous agent will tend to give the appropriate response to those characteristics. The possession of virtue moreover seems to be something that admits degrees of competence along a spectrum that goes from knowledge about what to do to a more complex one about why to do a certain thing. This element depends on two factors that concern our practice of attributing virtuous traits:

1) we sometimes attribute a virtue even when those who perform the corresponding action are not able to give an articulated justification of what they have done and/or their actions are not cross situationally consistent.
2) we believe that those who give an articulated justification of their virtuous actions, and tend to manifest the virtuous trait in a coherent plurality of situations, have a greater knowledge of virtue than those who do not. Unlike the former, these agents tend to know how to distinguish circumstances that require a virtuous response from similar ones that do not.

An entirely brave person, for example, will tend to discriminate between the real danger from a mere apparent threat and to avoid facing danger just for the sheer pleasure of the adrenaline that follows. It will also be able to recognize situations that require a courageous response and to be motivated accordingly in a variety of situations: not only those involving physical confrontation, but also those involving, for example, the defense of an unpopular idea or pursue a complex line of research never explored before. In a similar way, an entirely accurate person takes into account that research is both time- and resource-consuming and is able to put different investment policies in place depending on circumstances. These agents seem to possess greater knowledge of the virtue of those who simply act virtuously in a limited number of cases and lack the ability to formulate a justification for what they do. They are able to articulate the reasons in favour of virtuous behaviour in a plurality of contexts and use them to justify their conduct to themselves and others. For those who possess this knowledge virtue will not be a mere disposition to act, but a disposition to act infused by a reflective ability that involves the mastery of concepts. As with other character traits, finally, virtues are *profound qualities* that reveal what kind of person is the one who possesses them. The attribution of character traits is in fact a common practice that we use as much to get a general idea of someone we do not know well as, if we know him better, to predict what he will say or do or to explain why she has made certain choices in the past. Moving from the spectator's point of view to that of the agent, there are facts that are typically explained (retrospectively) by referring to motives that depend on general principles of our conduct that survive that individual case. Although not all of these principles are something we necessarily approve of, some of them are aspects of our character that are valued by ourselves and others and that we strive to maintain, creating opportunities to test them and reflect on how we express them in our behaviour. In these cases, virtue becomes a crucial element of our narrative and plays a normative role in our future choices.

In our research we are going to use two different types of virtues: the *intellectual* or *epistemic virtues* and the *virtues of character*. The former involve dispositions to exercise a set of capacities that relate to how we acquire our beliefs and communicate them to others. The latter, on the other hand, concern functional ways in which we enter into relationships with ourselves and interact with others by promoting our own and their good. The former are developed primarily through teaching by specialized staff (professors; research directors; etc.). The latter, on the other hand, require having entered into a non-emulative educational relationship with figures who have the role of educators in the community (parents, teachers, etc.).

Below we provide an open-ended list of the virtues that enable the constitution of a good research practice and constitute something that should be taken into account by those making comparative judgments about the relative value of different research practices in addition to existing and currently used individual bibliometric indicators. For an extensive discussion on these individual bibliometric indicators, see Schubert and Schubert (2019) and Wildgaard (2019). Schubert and Schubert (2019) offer an overview on h-index related indicators while Wildgaard (2019) presents a summary of existing author-level indicators of research production.

The virtues that enable the constitution of a good research practice include:

a) Intellectual virtues:

Accuracy: it is the disposition that consists in the care with which the individual researchers collect data that will constitute the pool of information shared in the research practice. Since collecting and evaluating information is time and resource consuming, the accurate researcher is one who implements several "policies of investigation" that are appropriate to the research circumstances (Williams, 2002).

Sincerity, Honesty: it is the disposition to tell the truth to others and, when this does not happen, it is the capacity to indicate good reasons why this did not happen where good refers to the fact that these reasons have a constitutive reference to the interests of other people (McIntyre 1985, Williams 2002).

Creativity: is the ability, which finds expression both in our social interactions with others and in the products of our research, to produce something that not only has value but is characterized by the elements of novelty and the capacity to arouse surprise in others (Swanton 2003, pp. 162, 165).

b) Virtues of character useful to oneself:

Humility: it is the ability to accept the authority of the standards related to the rules that define the practice. I have to recognize that other participants know rules and know how to apply them better than I do. I have to be willing to learn from these people and accept their criticism (MacIntyre 1985, p. 193).

Pride: it is evaluative attitudes towards ourselves (Ardal, 1966; Cohon, 2008; Taylor, 2015). Unlike other emotions, which simply motivate us to pursue or avoid objects, this traits of character fix our attention on persons, casting a positive or negative light on them. If I am proud of my child's success at school, my pride does not fix my attention on the 'merits of my child,' and still less on 'me in the role of father,' but on the whole of myself. As Cohon has rightly said, "when I feel pride, I am proud of something in particular [its cause] … But the attitude of pride is a pleasure or satisfaction not in that particular accomplishment or possession, but in myself in my entirety" (Cohon, 2008, p. 166). We believe that the pride associated with one's own achievements in research and the consequent approval of one's peers or superiors is a fundamental spring that drives researchers to perform at best in their area of research (Tangney, 1999).

Patience is the ability to curb one's own urge to complete a research in order to obtain as soon as possible the gratification of a positive result. To be able to wait and to be guided by a cautious skepticism that prompts us to control accurately the different steps of our investigation.

Prudence: is the capacity to sacrifice the satisfaction of less important pleasures closer in time than the satisfaction of more distant but more important pleasures. Where the degree of importance is defined with respect to the long-term objectives that characterize our lives (Parfit 1984).

Resilience: together with pride, this ability is indispensable to move forward in the research. It allows us to leave behind failures (paper rejected, unfunded projects, etc.) and to focus on future projects (Hormann, 2018).

c) Virtues of character useful to others

Courage is the capacity to risk damage or danger to oneself when individuals, values, goals that are crucial to the existence of the practice are at stake. Courage is therefore a way of showing that our attachment to these elements of the practice is genuine (MacIntyre, 1985, p. 192).

Empathy, Benevolence: in line with extensive literature, by this term we mean the human ability to feel the emotions and feelings of other people through a vicarious feeling that is similar to that of the person with whom we sympathize. We do not believe, however, that empathy in itself is a virtuous capacity in research practices. Since empathy is an instrument for reading the other's mind, it can also be used to manipulate other researchers in malicious ways. Empathy must be cultivated in such a way that it is rooted in the benevolent tendencies of human beings (Batson 2017, p. 2). In this way, empathy can allow the creation of a climate of *trust* between those who work within research institutions. Indeed, mutual trust is an indispensable component in these practices given the fundamental fact of the asymmetry of power that characterizes those interactions (Baier, 1991).

Integrity: is the willingness to behave in such a way that our actions are the outcome of our deepest values and commitments and that we tend to refuse making them hostages to imposed obligations or duties that we do not endorse on reflection.

Justice: following Aristotle (Aristotle 2014, Book V), we distinguish justice into two aspects. The first relates to the ability to promote the good of other researchers as members of our institution and research unit, that is, those to whom we are bound by ideal bonds and common respect for the rules that define our research community. The second is in turn divided into two spheres. The first concerns the ability to distribute the goods of research practice appropriately on the merits. The second, on the other hand, is the ability to compensate for or punish poor distributions or misconduct.

Practical wisdom: it is a kind of *super-virtue* essential to make each virtue effective. In line with Aristotle, we believe that this rational capacity enables the virtuous agent to acknowledge and respond properly to the items in the field of the research practice, choosing the appropriate means for their own ends (McDowell, 1979). Moreover, it also allows the different virtues within an individual's character to operate and develop harmoniously with each other.

A first attempt to develop a questionnaire for the evaluation of virtues in research practices has been done by Daraio and Vaccari (2020). Our Table 1 below, elaborating further Table 1 of Daraio and Vaccari (2020, p. 1067-1068), proposes some examples of questions to consider in evaluating the virtues of researchers.

Table 1. Examples of questions to include in the Questionnaire on Researchers' Virtues

a) Intellectual Virtues	**Questions**
Accuracy	- Do you thoroughly collect all the pieces of information that constitutes the body of knowledge on which the practice revolves? - Do you evaluate this disposition as instrumental or intrinsic? - Do you have a tendency to share data, results, methods, ideas, techniques, and tools used in the practice?
Sincerity, Honesty	- Do you think there are circumstances in which your colleagues can be manipulated? - Are you inclined to admit publicly when you make mistakes?
Creativity	- Are you able to explore and follow your own line of research? - Do you have the tendency to question your own ordinary experience and look with suspicion at what is the result of habit?
b) Virtues of character useful to oneself	**Questions**
Humility	- Are you incline to recognize that other researchers (participants to the practice) know rules and know how to apply them better than you do? - Are you to be willing to learn from these people and accept their criticism?
Pride	- Are you able to gain an impartial knowledge of one's own qualities? - Do you have the ability to feel fulfilment for academic success through demonstrating competence according to social standards and to draw strength from one's achievements? - Do you think you have a stable awareness of your own value that is not shaken by the successes of others? Do you have the capacity to enjoy and congratulate other people' accomplishments?

Patience	- Are you able to curb the rush to hastily complete a search to achieve the gratification that comes with a prima facie positive result? - Are you willing to be guided by a cautious scepticism that prompts you to control accurately the different steps of our investigation?
Prudence	- What do you think about people who tends to restraint of actions, inclinations, and impulses that are likely to upset others? - Do you tend to respect the fundamentals of the research practice within which one works?
Resilience	- Would you describe yourself as a person who leaves behind failures (paper rejected, unfunded projects, etc.) and to focus on future projects?
c) Virtues of character useful to others	**Questions**
Courage	- Are you willing to risk damage or danger to oneself when individuals, values, goals that are crucial to the practice are at stake? - Do you have a propensity to apply for highly competitive grants?
Empathy, Benevolence	- How do you feel about people who are sensitive to the suffering of their colleagues caused by failures or exclusions? - How do you feel about people who seeks feedback to improve interactions with others? - Do you think that preservation and promotion of the welfare of people with whom one is in frequent personal contact is important? - How would you describe collegial engagement?
Integrity	- Do you think that to hold beliefs that are consistent with actions has always a positive value?
Justice	- Does the practice of distributing funds, grants and awards according to merit always have a positive value? - Do you tend to think that the situations in which researchers work have commonalities that would allow for comparisons of their accomplishments?
Practical wisdom	- Would you describe yourself as someone who tends to choose the most effective means to achieve their ends? if not, why? - Have you ever been in a situation of conflict between things that have identical value to you? If so, what considerations do you turn to in an attempt to resolve the conflict?

Note: The content of this table is an elaboration of the content of Table 1 by Daraio and Vaccari (2020, pp. 1067-1068).

Designing an ontology for the modelling of researchers' virtues: Implementation aspects

Formally, an ontology in Description Logics is a knowledge base. It is a couple (pair) O=<TBox,ABox>, where TBox is the Terminological Box that represents the *intensional* level of the knowledge or the conceptual model of the portion of the reality of interest expressed in a formal way; and ABox is the Assertion Box that represents the *extensional* level of the knowledge or the concrete model of the portion of the reality expressed by means of assertions (instances).

The use of ontology-based modelling in our context allows us to implement *cognitive interviewing methodology* to address the challenges outlined in the previous section.

Cognitive interviewing is a psychologically oriented method for empirically studying the ways in which individuals mentally process and respond to survey questionnaires. Cognitive interviews can be conducted for the general purpose of enhancing the understanding of how respondents carry out the task of answering survey questions. However, the technique is more commonly conducted in an applied sense, for the purpose of pretesting questions and determining how they should be modified, prior to survey fielding, to make them more understandable or otherwise easier to answer. The notion that survey questions require thought on the part of respondents is not new and has long been a central premise of questionnaire design. However, cognitive interviewing formalizes this process and it has become an interdisciplinary field (for an overview, see Willis, 2004 and Miller et al., 2014).

An ontology-based semantic modelling approach offers several advantages, including:

i. a conceptual specification of the domain of interest, in terms of knowledge structures;

ii. the mapping of such knowledge structures to concrete data (the answers of the questionnaire);

iii. the reasoning over the abstract representation of the domain prior to the data collection;

iv. a flexible conceptual system that can be easily updated;

v. an open conceptual system that can be used as a common language for the research community.

The languages for representing ontologies and taxonomic knowledge, and the reasoning algorithms that can make inferences on the basis of such knowledge have been addressed by a large body of research in Artificial Intelligence and Knowledge Representation. Their formal characterisation is nowadays based on Description Logics (Baader et al., 2003), which provides a syntax for concept and role expressions and formal semantics to interpret them in a set theoretic framework. Concepts (i.e. classes) model sets of individuals and roles model binary relations. The representation of ontological knowledge is thus achieved by defining concepts and the properties (relations) that link them to other concepts in the domain of interest. The concepts are arranged in a hierarchical structure based on the subsumption relation (i.e. set containment). Along with the formal language, systems that allow to model and use ontologies are accompanied by various forms of syntactic representation, including graphical models. Protegé (Gašević et al., 2009) is a standard tool that builds its success, among other things, on the capability to handle multiple syntactic representations that allow the user to model the domain of interest using the most convenient notation, while grounding it to a well understood formal counterpart. Another key feature of Protegé is the decoupling of the representation from the reasoning tool that is adopted to make inferences. Protegé, for all the reasons explained above, will be used for the development of the ontology for the assessment of researchers' virtues.

Related works

While there is a rich literature and several approaches for extracting information through ontologies, for a review see Wimalasuriya and Dou (2010), the literature on ontological modelling in support of questionnaires development is scant. Notable exceptions include Sherimon et al. (2013 and 2014) where an ontology-based model for gathering the patient medical history based on dynamic questionnaire ontology is developed. The model is implemented and explained for diabetes domain by using Protegé. Another interesting contribution is Borodin and Zavyalova (2016), in which the authors focused on the problem of semantic representation of questionnaires. They constructed the generic ontological model of questionnaire, which provides a possibility of question structure description including complex questions with a set of answers of different kind and question order, including skipping and branching.

Surveys in fact, may be conducted to gather information through a printed questionnaire, over the telephone, by mail, in person or on the web, etc. and the structure of survey questionnaires and feedback if described in ontological terms provides an opportunity not only to structure the survey data but also to analyse the responses.

We contribute to enrich the limited existing literature, showing the potential of ontology based modelling for challenging and intriguing topics as the "good" evaluation of research practices, that relates the development of a questionnaire for the assessment of the motivations and the stable threats of characters (or virtues) of researchers.

Concluding remarks

This paper is realized within a challenging and interdisciplinary research project on "Evaluating Evaluations" with the financial support of Sapienza University of Rome. Based on the conceptual tools used by normative ethics, in particular from the perspective known as ethics of virtues, we have started to develop a questionnaire capable of revealing the presence and quantity of virtues in individual researchers. We believe that the use of ontology-based modelling in this context might enable us to further implement a cognitive interviewing methodology which may help us to address the many challenges of this research field.

Acknowledgments

The financial support of the Sapienza University of Rome (through the Sapienza Awards no. PH11715C8239C105 and no. RM11916B8853C925) is gratefully acknowledged.

References

Ardal, P. S. (1966). *Passion and Value in Hume's Treatise*. Edinburgh: Edinburgh University Press.

Aristotle (2000, revised edition 2014). *Nicomachean Ethics*, R. Crisp (Ed.), Cambridge, Cambridge University Press.

Baader, F., Calvanese, D., McGuinness, D., Patel-Schneider, P., Nardi, D. (2003) (Eds.). *The description logic handbook: Theory, implementation and applications*. Cambridge university press.

Baier, A. (1991). *A Progress of Sentiments: Reflections on Hume's Treatise*. Harvard: Harvard University Press.

Batson C D (2017) The Empathy-Altruism Hypothesis: What and So What? In Emma M. Seppälä, Emiliana Simon-Thomas, Stephanie L. Brown, Monica C. Worline, C. Daryl Cameron, and James R. Doty (eds.), *The Oxford Handbook of Compassion Science*, Oxford, Oxford University Press.

Borodin, A. V., & Zavyalova, Y. V. (2016) An ontology-based semantic design of the survey questionnaires. In 2016 19th Conference of Open Innovations Association (FRUCT) (pp. 10-15). IEEE.

Cohon, R. (2008). *Hume's Morality: Feeling and Fabrication*. Oxford: Oxford University Press.

Dahler-Larsen P. (2011), *The Evaluation Society*, Stanford Univ. Press, Stanford.

Daraio C. Vaccari A. (2020) Using normative ethics for building a good evaluation of research practices: towards the assessment of researcher's virtues, *Scientometrics,* 125, 1053-1075.

Gašević D.; Djurić D.; Devedžić V. (2009). *Model Driven Engineering and Ontology Development* (2nd ed.). Springer.

Hochschild, J. L. (2009) Conducting intensive interviews and elite interviews. In Workshop on interdisciplinary standards for systematic qualitative research. National Science Foundation.

Hormann, S. (2018). Exploring Resilience: in the Face of Trauma. *Humanistic Management Journal,* 3(1), 91–104.

Kvale, S. (2008) *Doing interviews*. Sage.

Larson, R. C., Berliner, L. (1983), On evaluating evaluations. *Policy Sciences*, 16(2), 147-163.

MacIntyre, A. (1981 first ed., 1985) *After Virtue*, London, Duckworth.

Murdoch, I. (1998) *Existentialists and Mystics: Writings on Philosophy and Literature*. Allen Lane/the Penguin Press.

McDowell, J. (1979). Virtue and Reason. *The Monist, 62*(3), 331–350.

Miller, K., Chepp, V., Willson, S., & Padilla, J. L. (Eds.). (2014). *Cognitive interviewing methodology*. John Wiley & Sons.

Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Rabionet, S. E. (2011) How I Learned to Design and Conduct Semi-Structured Interviews: An Ongoing and Continuous Journey. *Qualitative Report*, 16(2), 563-566.

Sherimon P.C., Vinu P.V., Krishnan R., Takroni Y., (2013) Developing a Survey Questionnaire Ontology for the Decision Support System in the Domain of Hypertension, IEEE South East Conference, April 4-7, Florida, U.S.

Sherimon, P. C., Vinu, P. V., Krishnan, R., Takroni, Y., AlKaabi, Y., & AlFars, Y. (2014) Adaptive questionnaire ontology in gathering patient medical history in diabetes domain. In Proceedings of the first international conference on advanced data and information engineering (DaEng-2013) (pp. 453-460). Springer, Singapore.

Schubert, A., & Schubert, G. (2019). All along the h-index-related literature: A guided tour. In *Springer Handbook of Science and Technology Indicators* (pp. 301-334). Springer, Cham.

Snow N. (2014) Virtue intelligence, unpublished paper presented for the conference "Can Virtue Be Measured?" held by the Jubilee Centre for Character & Value, 9th – 11th January 2014, https://www.jubileecentre.ac.uk/485/conferences/can-virtue-be-measured.

Snow, N. (2010) *Virtue as Social Intelligence: An Empirically Grounded Theory*, New York, Routledge.

Snow N. (2012) Notes Toward an Empirical Psychology of Virtue: Exploring the Personality Scaffolding of Virtue, in *Aristotelian Ethics in Contemporary Perspective*, ed. by J. Peters, New York, Routledge, pp. 130-144.

Swanton, C. (2003). *Virtue Ethics: A Pluralistic View*. Oxford: Clarendon Press.

Tangney, J. P. (1999). The self-conscious emotions: Shame, guilt, embarrassment and pride (pp. 541–568). In Tim Dalgleish & M. J. Powers (eds.), *Handbook of Cognition and Emotion*, New York, Wiley.

Taylor, J. (2015). *Reflecting Subjects: Passion, Sympathy, and Society in Hume's Philosophy*. Oxford: Oxford University Press.

Wildgaard, L. (2019). An overview of author-level indicators of research performance. In *Springer Handbook of Science and Technology Indicators*, 361-396.

Williams, B. (2002). *Truth and Truthfulness: An Essay in Genealogy*. Princeton: Princeton University Press.

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Sage publications.

Wimalasuriya, D. C., Dou, D. (2010) Ontology-based information extraction: An introduction and a survey of current approaches, *Journal of Information Science*, Volume: 36 issue: 3, page(s): 306-323.

Wolcott, H. F. (2008) *Writing up qualitative research*. Sage Publications.

Open access policies and mandates and their practical implementation in Spanish public universities.

Daniela De Filippo[1,2] and Jorge Mañana-Rodriguez[2]

[1] *dfilippo@bib.uc3m.es*
LEMI (Metric Information Studies Laboratory, Department of Library and Information Sciences, Carlos III University of Madrid, Calle Madrid 126, 28903, Madrid (Spain)

[2] *jmanana@pa.uc3m.es*
Research Institute for Higher Education and Science (INAECU) (UAM-UC3M), 28903 Getafe (Spain)

Abstract

This paper analyses the open access (OA) output of Spanish public universities from 2011 to 2020. A bibliometric analysis was carried out using Clarivate Analytics' Web of Science Core Collection, and a set of bibliometric analyses was performed to provide an evidence-based response to the research questions (What are the most significant initiatives for promoting open access at higher education institutions at both the national level and the institutional level? What open access contribution are Spanish public universities making? What types of open access are used the most? How does OA affect the impact of scientific production? What type of open access increases impact? Is there a relationship between funded research and open access publications of results?). The bibliometric analyses looked at four dimensions: activity indicators, OA types, impact indicators and funding. The paper concludes that the institutional measures and actions fostering OA identified in universities' policies and mandates bear a correspondence with the bibliometric data. Spanish public universities have higher rates of OA publications, with a clear preference for Green Published and DOAJ Gold publications. OA publications are systematically more cited than non-OA publications (with Green Accepted in the lead), and there is a strong relationship among visibility, funding and open access.

Introduction

The early adoption of OA mandates in organizations such as the US National Institutes of Health and the UK Higher Education Funding Council (Van Noorden, 2014) has evolved steadily, spreading through both national and supra-national organizations. In 2020 the Registry of Open Access Repository Mandates and Policies records over one thousand OA mandates. Of key importance at the supra-national level, the launch of Plan S by Coalition S (an initiative backed by the European Commission, the European Research Council and the national research agencies and funders from twelve European countries) has triggered an intense debate on its practical implications and contributed to setting the OA policies at the centre of European research policy. Also at the supra-national level, the European Union's Horizon 2020 research grant scheme includes an explicit OA mandate. OA mandates and policies show a correlation between their intensity and their effects (Gargouri et al., 2012), and the economic implications are often non-trivial (Quadery et al., 2019).

The practical implementation of OA mandates was researched in depth by Larivière and Sugimoto (2018), who analysed more than 1.3 million papers reporting publicly funded research by a selection of US, UK and Canadian organizations. The authors identified important differences in OA percentages and OA types between funding programmes, agencies and fields of knowledge. De-castro and Franck (2019) conducted an in-depth analysis of the European Commission's pilot initiative to fund article processing charges (APCs) associated with finished projects financed under the 7th Framework Programme. They concluded that the transference of OA funding policies to specific institutions might have a positive effect on the overall efficiency of publicly assumed APCs.

In Spain the evolution of the rules and regulations affecting OA publications has one of its important landmarks in the 2011 Science, Technology and Innovation Law. The law's preamble establishes a favourable position toward open access policies as one of its 'measures for a science of the 21st century'. Section 37 of the law states that the personnel whose research activity is primarily funded by the national budget must post a publicly accessible digital version of the final contents published or accepted for publication by no later than 12 months after acceptance. In line with legislation, the main public research-funding instrument in Spain, the State RDI (Research, Development and Innovation) Plan, requires the open access publication of the findings of publicly funded research. In addition, in another a major milestone in the evolution of OA policies in Spain, the Spanish Conference of University Rectors, acting on their universities' behalf, has released a public commitment to open science. This includes a declaration supporting measures to foster the implementation of open science and containing several points directly related to open access and its management by Spanish universities.

The correspondence between policies and mandates and their practical implementation is key to understanding the effectiveness of the former and the real dimension of the latter. A detailed analysis of existing policies and the volume of OA publications and their types is the first step toward understanding the relationship between them and establishing a baseline for further cohort analyses.

Background

Interest in analysing open access is evident in numerous scientific publications that approach the subject from different perspectives. On the one hand, there are extensive discussions about the advantages and limitations of open publishing (Kurtz & Brody, 2006; Beall, 2012) and analyses of its scope and implications (Suber, 2003, 2005; Zuccala, 2009). The relationship among open access, visibility and increased citations has also been explored (Harnad & Brody, 2004; Eysenbach, 2006; Moed, 2007; Gargouri et al., 2010; Suber, 2012; Science-Metrix, 2018). The relationship between impact and open access type is a topic of constant interest as well (Perianes & Olmeda, 2019), with some studies focusing on the Green route (Gargouri et al., 2010; Harnad & Brody, 2004; Moed, 2007; Gumpenberger, Ovalle-Perandones & Gorraiz, 2013), while others consider the Gold route more promising (Jubb et al., 2011; Torres-Salinas et al., 2018). Another endeavour of particular interest is to identify the proportion of publications in open access in specific databases (Björk et al., 2010; FECYT, 2016), disciplines and countries (Gaugouri et al., 2012; Archambault, 2013).

Some open access policies' impact has been investigated in studies analysing the effect of the mandates of international RDI funding agencies (European Commission, 2016b) and national funding agencies, as in the case of Canada (Zhang & Watson, 2017) and Spain (Borrego, 2015). More-recent studies have analysed the relationship among open access, DOI and dissemination on social media, showing how different access modalities can contribute to the visibility of research results (Laakso et al., 2017; Pinowar et al., 2018; Science-Metrix, 2018; Borrego, 2017).

Given that higher education institutions are responsible for a major share of a country's total publication output, several authors have studied open access in universities. Studies such as that of the universities of Sweden (Fathli, 2011), Norway (Elbaek, 2014) and the Netherlands (Bosman and Kramer, 2018) have shown the diversity of case studies between countries and disciplines. Another recent study (De Filippo and Mañana, 2020), which analyses open access in the universities belonging to the YERUN network, demonstrates the importance of promotion policies in universities and how such policies affect research visibility.

Study contexts and objectives

As mentioned, there is a particular interest in mainstreaming open access in higher education institutions. Several studies have focused on aspects related to certain variables (percentage of available publications, types of policies implemented, impact and citation). In this research, however, we consider open access to be very much related to aspects such as regulations and

mandates, access to funding, the national and institutional context in which research is done, impact and visibility.
The study focuses on the case of Spanish universities. These institutions are of interest because, as stated before, an explicit national policy exists about the mandatory publication of all findings of publicly funded projects. Furthermore, Spanish research centres and higher education institutions have demonstrated high levels of availability in open access. The YERUN network study mentioned before (De Filippo & Mañana, 2020) shows that the four Spanish universities in the network have open access percentages that are up to 20 percentage points higher than the country's average. Another study on centres of excellence in Catalonia shows that nearly 70% of production is available by some means (Rovira et al., 2019).
One of the studies of the particular case of Spanish universities is Melero et al. (2018), which examines the degree of OA archiving of 28 Spanish universities between 2012 and 2014, finding an asymmetric distribution of articles deposited in repositories. Meanwhile, the 'Open Access Observatory', an information system developed by 11 Catalan universities, analyses the evolution of open access in Catalonia and shows a constant growth from 2011 to 2019 (Observatori de l' accés obert, 2020).

The authors of this paper are currently participating in a research project on the implementation of open science and open access in Spanish universities. This research is funded by the Spanish National Science and Technology Plan and will continue until 2023. In the light of the evidence and the importance that the Spanish government has given to open science practices, the main objectives of this study are twofold:

- To identify and analyse policies and mandates for the promotion and development of open access at the level of the Spanish Public University System (SPUS).
- To analyse practices related to open access to scientific publications, considering their evolution and main characteristics.
- To propose a typology of universities considering open access practices and their relationship with funded research and visibility.

The following research questions have been established:

Q1. What are the most significant initiatives for promoting open access at higher education institutions at both the national level and the institutional level?
Q2. What is public universities' contribution to the total number of open access Spanish publications?
Q3. What types of open access are used the most?
Q4. How does OA affect the impact of scientific production? What type of open access increases impact?
Q5. Is there a relationship between funded research and open access publications of results?

Source and methodology

Source

Two types of information sources were used:

- *Spanish national regulations.* Regulations and official websites were consulted to find out about aspects related to open access in higher education institutions in Spain.
- *International multidisciplinary bibliographic and bibliometric databases*. The publications by Spanish public universities included in the Clarivate Analytics Web of Science's Core Collection (Science Citation Index, Social Science Citation Index and Arts and Humanities Citation Index) for 2011-2020 were retrieved and analysed.

Methodology

The study was carried out in two phases:

- *Spanish national regulations*

  The main sources for the analysis of the milestones and evolution of OA in Spain were the articles reviewed in this document and the publicly available reports of official Spanish organizations such as ministries and universities. The *Melibea* database, which contains information on institutional OA policies and repositories for a number of Spanish universities (Acceso Abierto, 2018), was also queried to systematize the evolution of OA policies at the institutional level.
- *Bibliometric analysis*

  In this phase all scientific publications from Spain and all scientific publications from each Spanish public university were identified (through the 'organization enhanced' search).
  Publications of all document types and languages from 2011 to 2020 were recovered.
  After the documents were downloaded and cleaned up, the following bibliometric indicators were obtained.

  - Spanish publications in WoS per year (total and OA), growth rates
  - SPUS contribution (total and OA) to Spanish publications, growth rates
  - Percentage of OA publications per university
  - Type of OA in Spain and in each university
  - Citations/document in OA publications versus non-OA publications
  - Citations/document by open access type
  - Percentage of highly cited papers (HCPs) in total publications and OA publications
  - Percentage of OA in funded publication and non-funded publication (considering information given in funding acknowledgement fields). Funded publications are those resulting from a publicly or privately funded research project.

Open access documents were analysed according to the five types of OA provided in the WoS database.

1. **DOAJ Gold:** Articles published in journals listed in the Directory of Open Access Journals (DOAJ) as Gold in accordance with their licence type.
2. **Other Gold:** Other Gold OA articles identified as such by Impact Story or Unpaywall Database (Our Research). Most of the journals publishing this type of OA are hybrid journals.
3. **Bronze:** Free-to-read public access articles available at the publishers' website for which the specific licensing details are not clear
4. **Green Published:** Final published versions of articles hosted on an institutional or subject-based repository.
5. **Green Accepted:** Accepted manuscripts hosted on a repository (peer reviewed, but possibly not fully copyedited)

Case study description

Spain's university system contains 83 institutions, 50 of which are public, and 33, private (Ministry of Education, 2020). The latter group includes 16 Catholic universities, which are governed by specific legislation. Under Spanish law the purpose of all universities, be they public or private, is to provide a public service: higher education. This includes research, teaching and study.

Studies of the Spanish university system show that differentiation between types of institutions clearly identifies the existence of public and private 'subsystems' with very different characteristics and activity patterns. Public universities clearly outnumber private universities in terms of their absolute number of institutions, professors and students (Grau Vidal, 2012). Although the number of private institutions has climbed rapidly, due to their much smaller size private universities account for 10% of the academic staff and 12% of the students. Their contribution to research is much smaller still: 4% of the total scientific output. As shown by Casani et al. (2013), public universities produced 96% of the papers published in Web of Science (WoS) and received 99% of the citations. These data are clear indications of the scale and quality of the public sector. Moreover, its research is more visible, for Spain's public universities publish a larger percentage of their papers in Q1 journals than do the country's private universities.

On the basis of this evidence, this study focuses on the Spanish Public University System. The list of institutions included and their acronyms are shown in Appendix I.

Results

Spanish national regulations

Some Spanish institutions can be qualified as 'early adopters', such as the Universitat Oberta de Catalunya (UOC) and the Universidad Politécnica de Madrid (UPM). These universities developed OA regulations prior to the development of the official OA mandate (the 2011 Science, Technology and Innovation Law). The vast majority of institutions rolled out their OA regulations within five years after the law was enacted, while other organizations, such as the Spanish National Research Council (CSIC) and the Rectors' Conference (CRUE), implemented their standards or declarations of support only after developments and OA mandates were consolidated in other organizations (Figure 1).

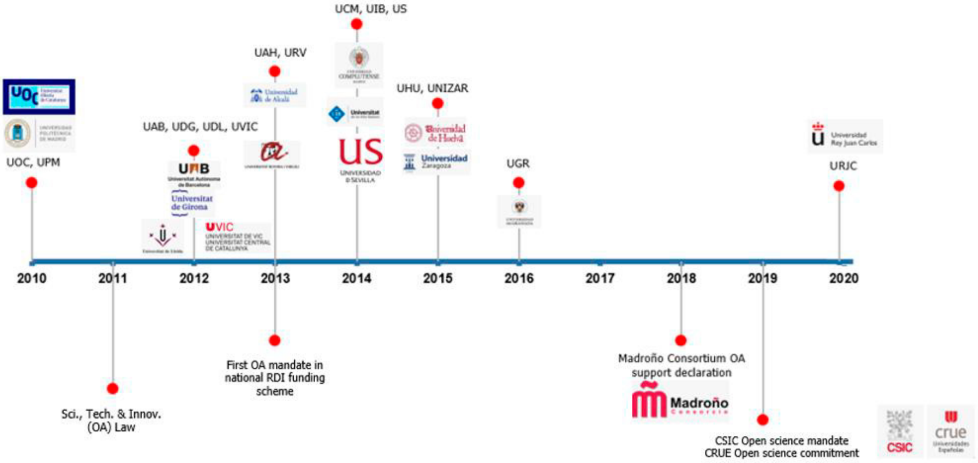

UCM, UIB, US
UAH, URV
UHU, UNIZAR
UAB, UDG, UDL, UVIC
UOC, UPM
UGR
URJC
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
First OA mandate in national RDI funding scheme
Sci., Tech. & Innov. (OA) Law
Madroño Consortium OA support declaration
Madroño
CSIC Open science mandate
CRUE Open science commitment
CSIC
crue

Figure 1: Timeline of implementation of OA legislation in Spanish universities (section above year line) and main policy milestones (section below year line)

For the specifics of the university regulations in place, we used the Melibea database (available at https://www.accesoabierto.net/politicas/). It is a directory and estimator of policies in favour of open access to scientific production developed by the 'Acceso abierto a la ciencia' (Open Access to Science) Research Group (CSIC and University of Barcelona). Melibea contains an extensive review of 31 repositories and the OA policies of Spanish universities. Going by the 2018 data, 51.6% of these repositories do not specify which version of the publication should be deposited, 41.9% call for the version corrected by the authors to be deposited, and 6.45% call for the published version to be deposited. On the subject of when to deposit publications, 38.7% require immediate deposit after publication and 35.5% do not provide specifications. Fifty-eight percent require the journal to have an OA policy prior to deposit, while 42% make of this feature a recommendation. The wide variability in the procedures and requisites for the deposit of research publications in institutional repositories is clearly an area for improvement; notwithstanding the organizational freedom universities hold under Spanish laws and regulations, common policies for institutional repositories are desirable, as they would increase the interoperability of the various platforms and foster national accountability for OA repository deposits.

While analysis of existing OA policies and developments does provide information on the evolution, extent and implementation of open access, it needs to be complemented with its quantitative counterpart. Accordingly, we retrieved and analysed universities' scientific output in Web of Science, which currently provides detailed data on the OA status of each contribution. The results of this analysis are detailed in the following section.

Bibliometric analysis

Between 2011 and 2020, Spain published 768 863 documents in the Web of Science Core Collection, 38.55% of them in open access. While the country's total production grew by 44%, open publications increased by 123%, increasing their share from 28.6% in 2011 to 44.8% in 2020. Higher education institutions produced more than 80% of Spain's publications in the period we analysed, making them the country's main producer of scientific publications. Open access documents at universities showed growth at a rate of 135%, which is higher than the Spanish average although the average number of publications in OA by universities (35%) is slightly lower than the country's average (Table 1).

Table 1. Annual evolution of publication (Web of Science, 2011-2020)

Indicator	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*	*Total*	*Growth rate*
Docs SPAIN	64194	68428	72007	72779	75078	77707	79989	83524	92457	82700	768863	44.03
Docs Spain OA	18339	21045	22433	25074	27890	32150	33945	37397	41052	37061	296386	123.85
% SPAIN OA	28.57	30.75	31.15	34.45	37.15	41.37	42.44	44.77	44.40	44.81	38.55	
Docs Univ	54489	58647	62462	63849	63302	69171	71679	75395	83804	76987	682785	53.80
Docs Univ OA	15988	18590	20121	22641	25261	29205	30963	34269	37610	34685	269333	135.24
% UNIV OA	24.91	27.17	27.94	31.11	33.65	37.58	38.71	41.03	40.68	41.94	35.03	

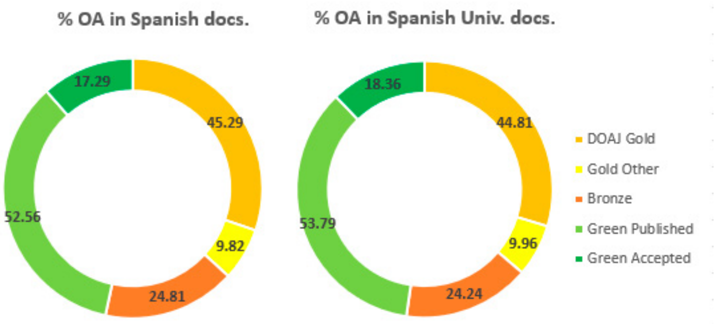


Figure 2. Distribution of publications by open access type (Web of Science, 2011-2020)

The most widely used types of access are Green Published and DOAJ Gold (Figure 2).
At the university level, the percentage of OA publication in the analysed decade ranges from 31% to 59%. Twenty-nine of the 50 public universities have percentages ranging between 30% and 40%, another 18 lie in the range from 40% to 50%, and only three exceed 50%. Looking at the most recent year only, the proportions increase significantly. Interestingly, open access rates are very high in HCPs, where three institutions publish over 80% of their output in OA (Figure 3).

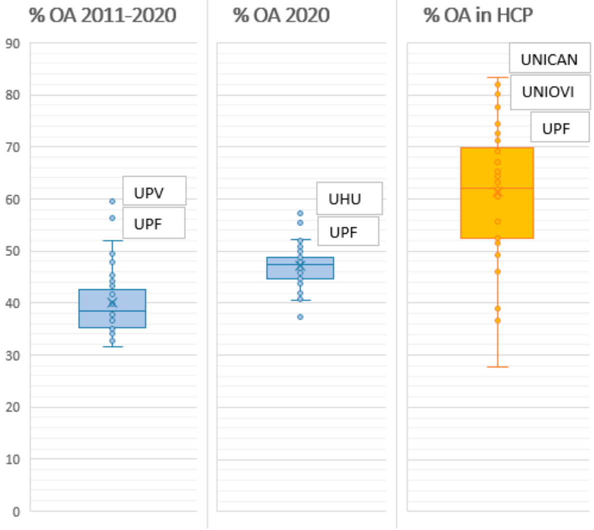


Figure 3. Distribution of open access (OA) publications at Spanish public universities

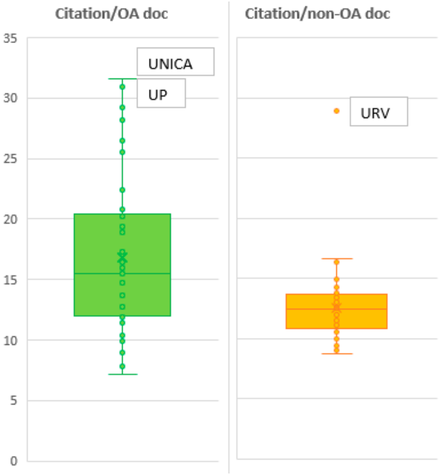


Figure 4. Distribution of impact (citations/doc) in OA and non-OA publications

Turning to the subject of publications' impact, open access documents received on average a higher number of citations; two universities have over 30 citations/document (Figure 4). We found that the two groups of publications (OA and non-OA) present statistically significant differences in their citations/document medians, U=639.5, Z=-3.755; p<.001, with a median of 15.51 for OA documents and a median of 12.41 for non-OA documents.

On comparison of open access documents' impact by access type, Green Accepted publications were found to receive on average more citations per document (Figure 5).

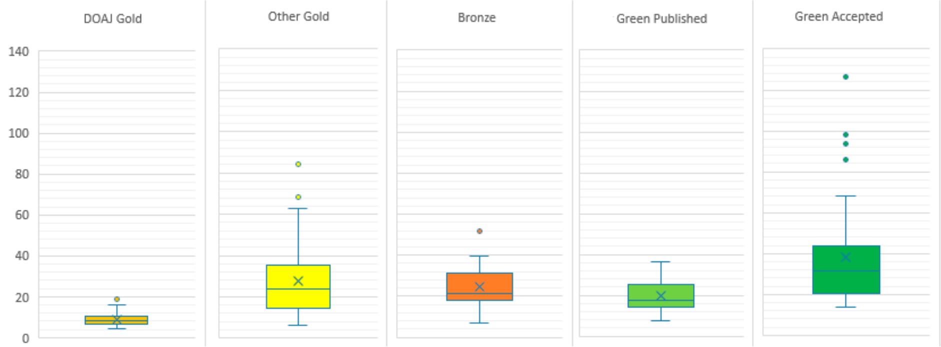


Figure 5. Distribution of impact (citations/doc) in publication by OA type

We also studied the percentages of OA publications in funded research and non-funded research. The two groups of publications (funded and non-funded) present statistically significant differences in their median percentages of OA publications, U=606.5, Z=-3.997; p<.001, with a median of 70% OA for funded documents and 65.22% OA for non-funded documents.

Discussion and conclusions

The landscape of OA rules in Spain is highly dynamic, with a wide range of directives codified into laws and regulations. The highest-ranked legislation addressing OA, the 2011 Science, Technology and Innovation Law, has not been followed up by specific regulations on the administrative consequences of noncompliance with its 37th article (which concerns open access). It has been implemented in a flexible framework, however, adapting its articles to the specific needs and requirements of various higher education, research-funding and research evaluation institutions.

The bibliometric analysis shows that the most widely used types of open access in Spain are Green and DOAJ Gold OA and that the growth rates for OA documents are greater for universities than for Spain as a whole. This could be partially explained by the existence of a wide variety of university regulations concerning OA publications and by the maintenance of a set of institutional repositories that provide researchers with the infrastructure required to deposit Green OA contributions.

The comparatively high percentages of OA documents among highly cited papers could be potentially explained by another variable: the non-independence between funding and OA percentages. Funding agencies' OA mandates seem to be an important factor, ceteris paribus, contributing to the statistically significant differences in OA proportions between funded research (which has higher OA proportions) and non-funded research. Although it cannot be deduced from the available data that HCPs are the result of funded research, if that were the case (which seems plausible if we assume that one of the expected results of most funded research is significant scientific impact), one potential explanation would be that HCPs are subject to funding entities' OA mandates. Self-selection bias, implying that most relevant papers of a given author are prioritized for OA publication, thus contributing to their 'citation advantage' is a plausible contributing factor in this case, although current data does not allow the authors calculating its relative contribution to high citation thresholds.

The citation advantage, an edge that has been extensively analysed in the bibliometric literature, is present in the case of the outputs studied. The differences in citations per document are statistically significant: OA documents receive a median of 15.51 citations per document, while non-OA receive 12.41. Within the various types of OA, Green Accepted presents the highest values. One factor that might help explain this observation is the existence of a variety of institutional repositories that could improve the accessibility of Green Accepted manuscripts.

Overall, the bibliometric data seem to be congruent with existing laws and regulations, although the specific paths that lead from official rules to specific, observable bibliometric phenomena are not completely clear. Further research might focus on such specific links between theory and praxis with the aim of identifying university types.

Acknowledgements

The results were found under 'Diagnosis of the OS in Spanish universities and instruments for its transformation and improvement (DOSSUET)', a project financed by the Spanish State Research Agency under the National RDI Plan's RETOS call.

This study was carried out with funds provided, also, by the Rámon y Cajal call (RYC2018-024894-I) from the Spanish Ministry of Science, Research and Universities.

References

Acceso Abierto (2018) *Melibea*. Available at: https://www.accesoabierto.net/politicas/

Archambault, E.; Amyot, D.; Deschamps, P.; Nicol, A.; Rebout, L.; Roberge, G. (2013). Proportion of open access peer-reviewed papers at the european and world levels: 2004–2011. *Sciencemetrix*. Retrieved from https ://www.scien ce-metri x.com/pdf/SM_EC_OA_Avail abili ty_2004-2011.pdf.

Beall, J. (2012). Predatory publishers are corrupting open access. *Nature*, 489(7415), 179-179.

Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., & Guðnason, G. (2010). Open access to the scientific journal literature: Situation 2009. *PLoS ONE*. https ://doi.org/10.1371/journ al.pone.00112 73.

BOE (2011). Law, O. 14/2011 on Science, Technology and Innovation (2011). *Boletín Oficial del Estado*, 131, 54387-54455.

Borrego, A. (2015). Measuring compliance with a spanish government open access mandate. *Journal of the Association for Information Science and Technology*. https ://doi.org/10.1002/asi.23422 .

Borrego, A. (2017). Institutional repositories versus researchgate: The depositing habits of Spanish researchers. *Learned publishing*. https ://doi.org/10.1002/leap.1099.

Bosman, J.; Kramer, B. (2018). Open access levels: A quantitative exploration using web of science and oaDOI data. *PeerJ Preprints*. https ://doi.org/10.7287/peerj .prepr ints.3520v 1.

Casani, F; De Filippo, D; García-Zorita, C; Sanz-Casado, E. (2013) Public versus private universities: assessment of research performance; case study of the Spanish university system. *Research Evaluation*, 23 (1): 48-61

De-Castro, P.; Franck, G. (2019). Funding APCs from the research funder's seat: Findings from the EC FP7 Post-Grant Open Access Pilot. *El profesional de la información*, 28(4).

De Filippo, D.; Mañana-Rodríguez, J. (2020). Open access initiatives in European universities: analysis of their implementation and the visibility of publications in the YERUN network. *Scientometrics*, 1-28.

Elbaek, MK. (2014). Danish open access barometer: Mapping open access to Danish research and creation of an online prototype for automated open access monitoring. *Sciecom Info*. Retrieved from https ://journ als.lub.lu.se/index .php/sciec ominf o/artic le/view/10238 /8629.

European Commission. (2016). *Ex-post evaluation of the seventh framework programme*. Retrieved from https://ec.europ a.eu/resea rch/evalu ation s/pdf/archi ve/fp7-ex-post_evalu ation /staff working_docum ent_annexes_part_2__en_autre docum ent_trava il_servi ce.pdf#view=fit&pagemode=none.

Fathli, M.; Lunden, T.; Sjogårde P. (2011). The share of open access in Sweden 2011: Analyzing the OA outcome from Swedish universities. *Sciecom Info*. Retrieved from https ://kth.diva-portal.org/smash /get/diva2 :78897 9/FULLT EXT01 .pdf.

Fundación Española para la Ciencia y la Tecnologia. (2016). *Informe de la comision de seguimiento sobre el grado de cumplimiento del articulo 37 de la Ley de la Ciencia*. Retrieved from https ://www.recol ecta.fecyt.es/sites /defau lt/files /conte nido/docum entos /Cumpl imien toOA.pdf.

Gargouri, Y.; Hajjem, C.; Larivière, V.; Gingras, Y.; Carr, L.; Brody, T., et al. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS ONE, 5*(10), e13636. https ://doi.org/10.1371/journ al.pone.00136 36.

Gargouri, Y.; Lariviere, V.; Gingras, Y.; Carr, L.; Harnad, S. (2012). Green and gold open access percentages and growth, by discipline. In *Conference presented at the 17th international conference on Science and technology indicators (STI)*. Retrieved from https ://eprin ts.soton .ac.uk/34029 4/.

Grau Vidal, F. (2012) *La Universidad pública española. Retos y prioridades en el marco de la crisis del primer decenio del siglo XXI*. Tarragona: Universidad Rovira Virgili Ediciones.

Gumpenberger, C.; Ovalle-Perandones, M.A.; Gorraiz, J. (2013). On the impact of gold open access journals. *Scientometrics, 96*(1), 221–238. https ://doi.org/10.1007/s1119 2-012-0902-7.

Harnad, S.; Brody, T. (2004). Comparing the impact of open access (OA) versus non-OA articles in the same journals. *D-Lib Magazine*, 10(6). Retrieved from https ://eprin ts.soton .ac.uk/26020 7.

Jubb, M.; Cook, J.; Hulls, D.; Jones, D.; Ware, M. (2011). Costs, risks and benefits in improving access to journal articles. *Learned Publishing,* 24(4), 247–260. https ://doi.org/10.1087/20110 402.

Kurtz, M.; Brody, T. (2006) The impact loss to authors and research. In, Jacobs, Neil (ed.) *Open Access: Key strategic, technical and economic aspects.* Oxford, UK. Chandos Publishing

Laakso, M.; Lindman, J.; Schen, C.; Nyman, L.; Bjork, B.C. (2017). Research output availability on academic social networks: Implications for stakeholders in academic publishing. *Electron Markets*. https ://doi.org/10.1007/s1252 5-016-0242-1.

Larivière, V.; Sugimoto, C. R. (2018). Do authors comply with mandates for open access?. *Nature*, 562(7728), 483-486.

Melero, R.; Melero-Fuentes, D.; Rodríguez-Gairín, J. M. (2018). Monitoring compliance with governmental and institutional open access policies across Spanish universities. *El profesional de la información*, 27 (4): 1699-2407

Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science and Technology,* 58(13), 2047–2054. https ://doi.org/10.1002/asi.20663.

Observatori de l'accés obert (2019). Available at: https://bibliotecnica.upc.edu/observatori?lang=es#presentacio

Perianes, A.; Olmeda, C. (2019). Efects of journal choice on the visibility of scientific publications: a comparison between subscription-based and full open access models. *Scientometrics,* 12*1*, 1737–1752. https ://doi.org/10.1007/s1119 2-019-03265 -y.

Piwowar, H.; Priem, J.; Larivière, V.; Alperin, J. P.; Matthias, L.; Norlander, B., et al. (2018). The state of OA: A largescale analysis of the prevalence and impact of open access articles. *PeerJ*. https ://doi.org/10.7717/peerj .4375. PMID: 29456894.

Quaderi, N.; Hardcastle, J.; Petrou, C.; Szomszor, M. (2019). *The Plan S footprint: Implications for the scholarly publishing landscape*. ISI (Institute for Scientific Information).

Rovira, A., Urbano, C., & Abadal, E. (2019). Open access availability of Catalonia research output: Case análisis of the CERCA institution. *PLoS ONE,* 14(5), e0216597. https ://doi.org/10.1371/journ al.pone.02165 97.

Science-Metrix. (2018). *Analytical support for bibliometrics indicators: Open access availability of scientific publications*. Montreal: Science-Metrix. Retrieved from: https ://www.scien ce-metrix.com/sites /defau lt/files /scien ce-metri x/publi catio ns/scien ce-metri x_open_acces s_avail ability_scien tific publi catio ns_report.pdf.

Suber, P. (2003). Removing the barriers to research: An introduction to open access for librarians. In *college and research libraries news*, 64. Retrieved from https ://eprin ts.rclis .org/4616.

Suber, P. (2005). Open access, impact, and demand: Why some authors self archive their articles. *BMJ: British Medical Journal,* 330 (7500), 1097.

Suber, P. (2012). *Open access*. Cambridge: MIT Press, 978-0-262-51763-8.

Torres-Salinas, D.; Robinson-Garcia, N.; Moed, H. F. (2019). Disentangling Gold Open Access. In *Springer Handbook of Science and Technology Indicators* (pp. 129-144). Springer, Cham.

Van Noorden (2014). Funders punish open-access dodgers. Nature 508:161. https://doi.org/10.1038/508161a

Zhang, L.; Watson, E. M. (2017). Measuring the impact of gold and green open access. *The journal of academic librarianship,* 43(4), 337–345.

Zuccala, A. (2009). The lay person and open access. *Annual Review of Information Science and Technology,43*(1), 1–62. https ://doi.org/10.1002/aris.2009.14404 30115

Appendix 1. Universities analysed

Acronym	*University*
EHU	Universidad del País Vasco
UA	Universidad de Alicante
UAB	Universidad Autónoma de Barcelona
UAH	Universidad Alcalá de Henares
UAL	Universidad de Almería
UAM	Universidad Autónoma de Madrid
UB	Universidad de Barcelona
UBU	Universidad de Burgos
UC3M	Universidad Carlos III de Madrid
UCA	Universidad de Cádiz
UCLM	Universidad de Castilla-La Mancha
UCM	Universidad Complutense de Madrid
UCO	Universidad de Córdoba
UDC	Universidad de A Coruña
UDG	Universidad de Girona
UDL	Universidad de Lleida
UGR	Universidad de Granada
UHU	Universidad de Huelva
UIA	Universidad Internacional de Andalucía
UIB	Universidad de las Illes Balears
UIMP	Universidad Internacional Menéndez Pelayo
UJAEN	Universidad de Jaén
UJI	Universidad Jaume I de Castellón
ULL	Universidad de La Laguna
ULPGC	Universidad de Las Palmas de Gran Canaria
UM	Universidad de Murcia
UMA	Universidad de Málaga
UMH	Universidad Miguel Hernández de Elche
UNAVARRA	Universidad Pública de Navarra
UNED	Universidad Nacional de Educación a Distancia
UNEX	Universidad de Extremadura
UNICAN	Universidad de Cantabria
UNILEON	Universidad de León
UNIOVI	Universidad de Oviedo
UNIRIOJA	Universidad de la Rioja
UNIZAR	Universidad de Zaragoza
UPC	Universidad Politécnica de Catalunya
UPCT	Universidad Politécnica de Cartagena
UPF	Universidad Pompeu Fabra
UPM	Universidad Politécnica de Madrid
UPO	Universidad Pablo de Olavide
UPV	Universidad Politécnica de Valencia
URJC	Universidad Rey Juan Carlos
URV	Universidad Rovira i Virgili
US	Universidad de Sevilla
USAL	Universidad de Salamanca
USC	Universidad de Santiago de Compostela
UV	Universidad de Valencia
UVA	Universidad de Valladolid
UVIGO	Universidad de Vigo

Excellence, Interdisciplinarity and Collaboration in Research Networks. Evidence from the Evaluation of the Special Research Programme (SFB) of the Austrian Science Fund (FWF)

Michael Dinges[1], Barbara Heller-Schuh[1], Robert Hawlik[1], Thomas Scherngell[1],
Anna Wang[1], Bart Thijs[2] and Wolfgang Glänzel[2]

[1] *michael.dinges@ait.ac.at*
AIT Austrian Institute of Technology GmbH, Giefinggasse 4, A-1210 Vienna (Austria)

[2] *wolfgang.glanzel@kuleuven.be*
Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven (Belgium)

Abstract

Excellent research is increasingly conducted within a complex web of interacting researchers, coming from different scientific and institutional backgrounds. The SFB Programme of the Austrian Science Fund aims to strengthen Austria as a location for high-level scientific research by supporting the establishment of sustainable, extremely productive and internationally visible research units. A theory-based evaluation was conducted following a mixed-methods approach to empirically examine the effects of the intervention on research excellence, patterns of collaboration and interdisciplinarity. The performed bibliometric analysis clearly shows that SFB succeeded in supporting exceptional research as evidenced by the outstanding publication and citation record of the funded projects. Moreover, the study reaffirms the SFBs' highly successful publication strategy as publications generally appear in high impact journals (with respect to their field) and receive more citations than expected for these journals. However, excellence in scientific outcomes is not paralleled by achievements in terms of interdisciplinarity and international collaboration. While the qualitative analysis emphasises the programme's perceived unique orientation towards promoting interdisciplinarity, interdisciplinarity is typically a result of very closely related (sub-) disciplines working in a sub-project. The evaluation resulted in a differentiated picture of the factors contributing to the objective of funding outstanding research.

Introduction

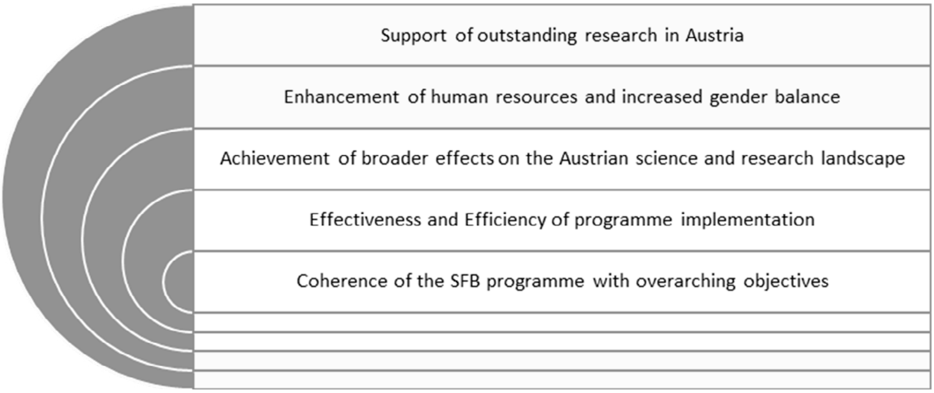
Particularly excellent research is increasingly conducted within a complex web of interacting researchers, coming from different scientific and institutional backgrounds (see, e.g., Powell and Grodal 2005, Scherngell 2013, among others). The growth of research collaboration as a phenomenon within the research system has attracted quite some attention in the academic literature but is also of major concern for research policymakers and funding organisations (see Scherngell 2013 for an overview). The potential benefits, costs and factors influencing the performance of research collaboration have been put under investigation, as research policymakers seek to design funding programmes that contribute to strengthening the competitiveness of national research and innovation systems for the benefits of society.

The SFB programme is the only network-based basic research programme in Austria. SFB aims to strengthen Austria as a location for high-level scientific research and enhance the competitiveness of the country's research and innovation system by supporting the establishment of sustainable, extremely productive and internationally visible research units, which are generally multi- or interdisciplinary and support institutions in developing strategic focus areas. Research networking programmes like the SFBs pick up these scientific insights and lay the foundation for nationally-funded research groups to address major scientific issues, to advance the frontiers of existing research and potentially pave the way for social and economic innovations. They seek to enhance 'research collaboration', which is a special form of collaboration, undertaken for research, where 'research' is implicitly seen as scientific research (Amabile et al., 2001, Katz & Martin, 1997).

In this article, we examine empirically the effects of the intervention on research excellence, patterns of collaboration and interdisciplinarity. Specifically, we address the question of identifying and measuring the contribution of the SFB Programme on outstanding research performance and the multi-/interdisciplinary character of the research networks. Thereby, we show that a mixed-method approach in evaluations, based upon a delineation of impact pathways allows combining different perspectives and increases research integrity.

Methodology

The article rests upon the findings of a comprehensive impact evaluation of the SFB programme, which was performed from 02/2019 until 02/2020 (Dinges et al. 2020) considering five evaluation dimensions (see Figure 1). The overall methodological approach to this evaluation study followed the principles of *theory-based evaluation* (Blamey and Mackenzie 2007, Carvalho and White 2004, Chen 1990, Mayne 2001) and *contribution analysis* (Mayne 2011, Mayne 2008). For gathering the relevant evidence, the study was informed by a mix of sources and recognised research methods (*mixed methods approach*).


Support of outstanding research in Austria
Enhancement of human resources and increased gender balance
Achievement of broader effects on the Austrian science and research landscape
Effectiveness and Efficiency of programme implementation
Coherence of the SFB programme with overarching objectives

Source: Dinges et al. (2020)

Figure 1: Main evaluation dimensions of the SFB Programme evaluation

The quantitative part of the analysis is based on various *bibliometric indicators*. In this programme-related study we were faced with special challenges linked to the specific task and the resulting research questions: In the first place, we mention the usual difficulty of attributing scientific publications correctly to the programme's outputs and impacts, and of comparing the results with those of individual projects and the national standards in the supported areas. After the appropriate framework for this was found, the challenges of using bibliometrics in this context had to be met. In particular, the evaluation needed to define how to identify and measure excellence, how to evaluate (international) collaboration and the extent of interdisciplinarity in the supported research since the programme aims to support outstanding research by creating long-term multi-/interdisciplinary and exceptionally productive research networks. Along with the data obtained from the SFB programme, we used bibliographic data on published results, control groups and the benchmark populations indexed in Clarivate Analytics Web of Science Core Collection (WoS).

The publication-activity and citation-impact study is based on bibliographic data and metadata extracted for WoS-indexed journal articles based on matched publication lists of the projects and on information in the funding acknowledgement section of the papers. Record matching techniques have been used to link the publication list with the bibliographic database including automated procedures and manual validation and reliability checks.

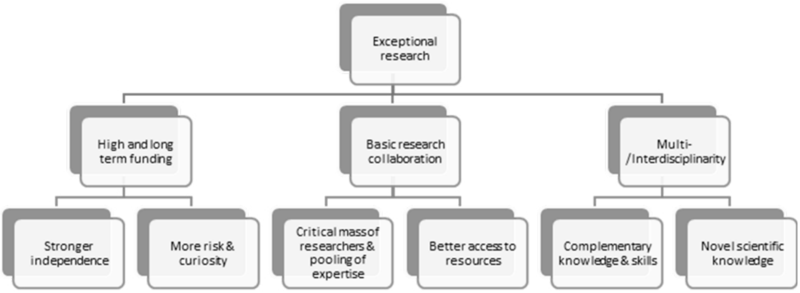
The publication-activity analysis is conducted for the entire period 2004–2017 and three subsequent and disjoint four-year sub-periods. The citation analysis is based on three-year citation windows (publication year and the two subsequent years) and, because of data availability, limited to the period 2004–2016. Only 'citable' papers (article, letter, note and review) have been taken into account. Bibliographic data are cleaned and processed to bibliometric indicators according to the standard rules in the field (see, e.g., Glänzel et al., 2009). A strict full-counting scheme has been applied to publications and citations. This applies to both subject matter and affiliation. The assignment of papers is based on the ECOOM classification system with science fields and disciplines. In this system, journals are assigned to and grouped into cognitive-logical disciplines. This scheme with 16 fields and 74 disciplines is hierarchically built on top of the Web of Science Subject Categories comprising about 250 categories (cf. Glänzel et al., 2016).

Propensity score matching was applied to identify an appropriate set of Stand-alone projects, another funding instrument of the Austrian Science Fund, as a *control group*. A *counterfactual analysis* was applied to identify statistically robust differences between network projects and stand-alone projects.

Finally, the qualitative analysis comprised *two online targeted surveys* including (sub)-project *coordinators* as well as *unsuccessful applicants*. The SFB (sub)-project coordinator survey yielded in the response of 211 (sub)-project coordinators of SFB networks out of 674 contacted, i.e. a response rate of 31.3%. *Individual interviews and two focus group interviews* with a total number of 31 participants were carried out with representatives from the networks, FWF, university management and R&I policy actors. Focus group 1 focused on researchers in the Social Sciences and Humanities (SSH), where eight carefully selected participants discussed their experiences with the SFB programme. Focus group 2 was dedicated to Medical and the Natural & Technical Sciences. This disciplinary approach enabled the validation of findings and additional data collection while taking into consideration the discipline-specific needs and requirements of researchers for excellent science and research funding.

The impact of SFB funding on outstanding research

SFBs incorporate a programme design, which inherently aims at exploiting the benefits from research collaborations. Through its programme characteristics, the programme may lead to *higher scientific quality, productivity and ground-breaking research results.*


Exceptional research
High and long term funding
Basic research collaboration
Multi-/Interdisciplinarity
Stronger independence
More risk & curiosity
Critical mass of researchers & pooling of expertise
Better access to resources
Complementary knowledge & skills
Novel scientific knowledge

Source: Dinges et al. (2020)

Figure 2: Impact pathways for funding outstanding research

Key hypotheses reflected in Figure 2 are that: 1) High and long-term funding of a group of researchers lead to stronger independence of researchers, more out of the box and curiosity-driven research; 2) The collaborative character of research facilitates the creation of a research-network of critical mass researchers, pooling of expertise, and better access to complementary resources; 3) The multi-/interdisciplinary character of the research networks allows researchers from different disciplines to look at similar questions, which may lead to entirely new combinations of complementary skills and knowledge and novel scientific knowledge.

Excellence: Publication activity and citation impact

In the bibliometric part of the analysis, we have used two basic sets of indicators, the first one reflecting publication activity and international collaboration, the second one regarding citation impact. The second set, in turn, consists of two parts, the standard set of relative indicators and the second part aiming to capture excellence as measured by outstanding impact. While the standard set of citation indicators sheds light on different aspects of factual citation impact and its different expectations, the second one supplements the linearly structured indicator model by performance profiles with seamless integration of measures of outstanding performance into the standard tools represented by the first set. This is summarised in nutshell below.

The citation-impact indicators form a triplet of mean observed, relative and normalised mean citations rates (MOCR, NMCR and RCR, respectively) according to Glänzel et al. (2009). This triplet allows the interpretation of how the observed citation impact (MOCR) relates to the expectations based on journals (RCR), where the papers in question have been published, and the disciplines (NMCR) to which these belong to. The most ‘favourable’ situation is NMCR>RCR>1, which means that the papers are published, on an average, in journals with a higher-than-discipline standard and receive even more citations than the standard set by the journals in which the papers are published. RCR<1<NMCR means that the latter standard is not reached and, for instance, NMCR<1<RCR means that the papers achieved a higher citation impact than expected based on the journals in they are published but these journals, on average, do not belong to the top journals in their discipline.

This set is extended by the citation classes according to the method of Characteristic Scores and Scales (CSS), which are compatible with the above set of relative indicators (cf. Glänzel et al., 2019). To capture excellence reflected by citation-impact, we apply four classes, which stand for ‘poorly cited’ (CSS1), ‘fairly cited’ (CSS2), ‘remarkably cited’ (CSS3) and ‘outstandingly cited’ (CSS4). Although CSS is not directly linked to percentiles, the distribution of papers over classes is about 70% (CSS1), 21% (CSS2), 6%–7% (CSS3) and 2%–3% (CSS4). Deviations of a profile from the reference standard provide a multifaceted picture of citation impact. All above-mentioned indicators can be calculated for any subsets such as international co-publications or research with a strong interdisciplinary orientation.

Bibliometric indicators for the SFB programme are provided in comparison with those for the matched set of FWF Stand-alone project publications and the complete Austrian population in Tables 1 to 3. The factual citation impact of SFB papers is above the journal expectation and that, in turn, is distinctly above the field-based expectation. The situation is stable, even improving. In other words, we find the most favourable situation here, NMCR>RCR>1, for the three periods between 2004 and 2016. Thus, the PIs of SFB projects in general publish in high impact journals (with respect to their field) and receive more citations than expected for these journals.

These indicators underline that the SFB programme clearly outscores both the Stand-alone projects and the Austrian overall publication set across all citation rates. The large shares of upper CSS classes reinforce the assessment of strong support for outstanding research by the programme.

The share of international papers in all SFB publications is increasing over time but is slightly below Austria's average share over the 14-year period. The interpretation of this deviation is not straightforward as co-publication patterns are influenced by multiple factors (subject profiles, size of community, expertise, economic factors, and others). Different selection biases might play a distinct role in the high-level SFB projects. International co-authorship proved to have a further positive effect on citation impact.

Table 1: Citation-impact indicators for the SFB programme publications in 2004–2016

Period	Publication Set	Publications	Share in total	Citations	MOCR	RCR	NMCR	NMCR/RCR	CSS1	CSS2	CSS3	CSS4
2004–2007	All	1,137		12849	11.30	1.26	1.88	1.49	47.1%	33.2%	12.5%	7.3%
	Int Collab	581	51.1%	7303	12.57	1.28	2.04	1.60	44.4%	31.8%	14.6%	9.1%
2008–2012	All	2,042		25434	12.46	1.15	1.91	1.66	43.2%	34.5%	15.9%	6.4%
	Int Collab	1,284	62.9%	17477	13.61	1.20	2.09	1.74	37.9%	37.2%	17.6%	7.3%
2013–2016	All	2,205		31043	14.08	1.19	2.07	1.74	41.7%	37.8%	14.1%	6.3%
	Int Collab	1,409	63.9%	23074	16.38	1.24	2.37	1.91	37.8%	38.2%	16.0%	8.0%

Source: Dinges et al. (2020) based on Web of Science Core Collection

Table 2: Citation-impact indicators for the FWF Stand-alone project publications in 2004–2016

Period	Publication Set	Publications	Share in total	Citations	MOCR	RCR	NMCR	NMCR/RCR	CSS1	CSS2	CSS3	CSS4
2004–2007	All	252		1,935	7.68	0.96	1.26	1.31	53.6%	32.1%	9.1%	5.2%
	Int Collab	144	57.10%	1,225	8.51	1.01	1.37	1.35	53.5%	29.2%	11.1%	6.3%
2008–2012	All	1,562		13,039	8.35	1.04	1.39	1.34	55.0%	30.6%	10.0%	4.4%
	Int Collab	950	60.80%	9,201	9.69	1.09	1.56	1.43	51.3%	31.5%	12.2%	5.1%
2013–2016	All	1,987		20,406	10.27	1.12	1.62	1.44	51.1%	31.6%	12.7%	4.6%
	Int Collab	1,229	61.90%	14,418	11.73	1.16	1.83	1.57	47.8%	31.4%	15.3%	5.5%

Source: Dinges et al. (2020) based on Web of Science Core Collection

Table 3: Citation-impact indicators for the Austrian publications in 2004–2016

Period	Publication Set	Publications	Share in total	Citations	MOCR	RCR	NMCR	NMCR/RCR	CSS1	CSS2	CSS3	CSS4
2004-2007	All	37,822		243,684	6.44	1.17	1.27	1.09	62.7%	25.2%	8.1%	3.9%
	Int Collab	20,445	54.1%	166,701	8.15	1.29	1.53	1.19	55.7%	28.6%	10.3%	5.5%
2008-2012	All	62,206		463,395	7.45	1.20	1.40	1.17	61.4%	25.4%	8.8%	4.4%
	Int Collab	38,434	61.8%	359,620	9.36	1.32	1.67	1.27	55.0%	28.3%	10.7%	6.0%
2013-2016	All	62,276		533,566	8.57	1.22	1.47	1.20	60.2%	26.3%	9.0%	4.5%
	Int Collab	42,419	68.1%	435,963	10.28	1.31	1.70	1.29	55.1%	28.4%	10.7%	5.8%

Source: Dinges et al. (2020) based on Web of Science Core Collection

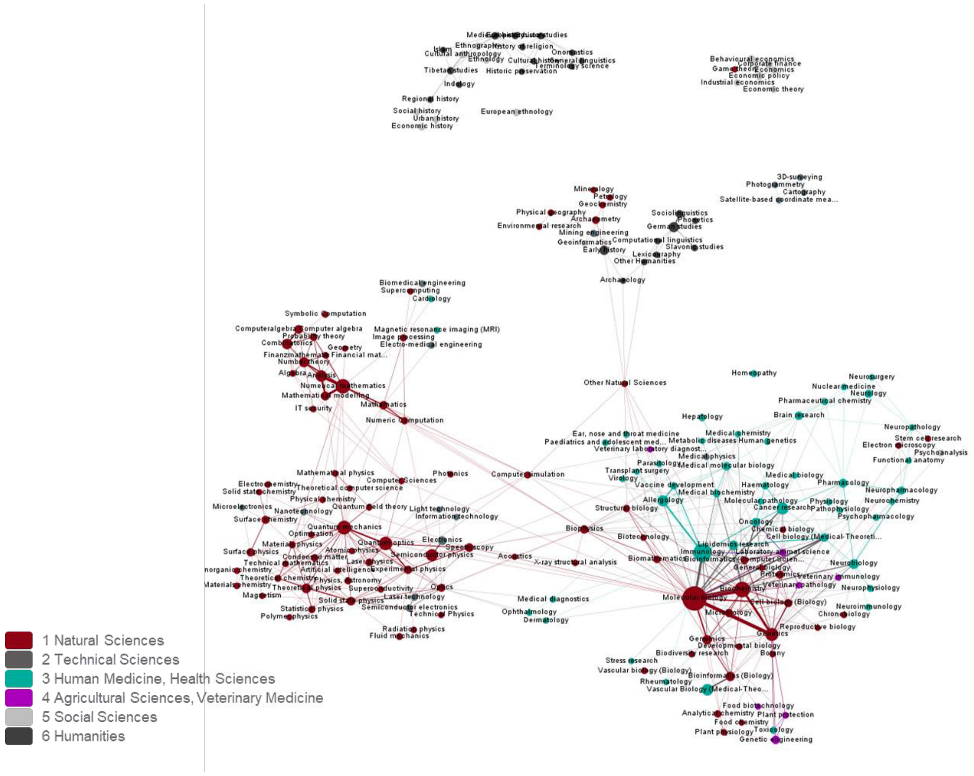
Survey and interview results highlight the perception that the SFB programme is primarily an instrument for highly experienced and established researchers. A large part of this pattern is attributed to the selection process and criteria, that place a strong emphasis on individual scientific excellence and publication track records, where young PIs with less experience and smaller networks are disadvantaged compared to senior researchers boasting scientific fame and awards.

Interdisciplinarity

As mentioned in the introductory section, SFB is laying great stress on supporting interdisciplinarity in research. Therefore, interdisciplinarity was analysed from three different angles.

1. Classification of sub-projects,
2. Publication results, and
3. Survey and interviews.

Ad 1. Interdisciplinarity was measured by the extent to which different disciplines of the three-level hierarchical ÖFOS 2012 classification are addressed within the same SFB sub-project. Based on SFB monitoring data of 331 sub-projects assigned to 195 sub-disciplines, we calculated and visualised the findings as a network, where nodes represent the sub-disciplines, which are connected when they are mentioned in the same project. The size of nodes represents the number of sub-projects, which are assigned to the respective subdiscipline. The size of the edges displays the number of joint projects between two sub-disciplines and the colour of the nodes refers to the highest level of the Austrian classification of science disciplines.



Source: Dinges et al. (2020) based on Web of Science Core Collection

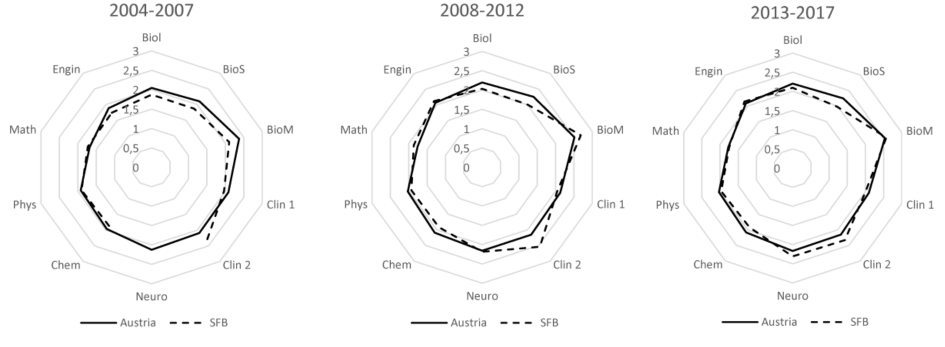
Figure 3: Connectedness of SFB sub-disciplines

The resulting network (Figure 3) displays four main clusters; among them two clusters with a strong interdisciplinary orientation and two disciplinary clusters.

- The first interdisciplinary cluster is the largest one in the network and situated in the bottom right area of the network. It comprises fields with clearly overlapping thematic orientation and is based on sub-projects assigned to “Biology”, “Medical-Theoretical Sciences, Pharmacy” and “Clinical Medicine”, especially “Molecular biology” (24.3 projects), “Immunology” (16.1 projects), and “Biochemistry” (12.7 projects).
- The second interdisciplinary cluster, located on top of the network, consists of several separate components and is mainly dedicated to the Social Sciences and Humanities with some connections to the Natural Sciences. In this cluster, no sub-discipline obtains a central position or is assigned to more sub-projects than others.
- The third cluster focusing mainly on “Physics, Astronomy” with single connections to disciplines in the Technical and Medical Sciences is positioned in the bottom left area of the network. Central disciplines are “Quantum mechanics” (10.6 projects) and “Quantum optics” (9.8 projects).
- The fourth cluster, finally, is formed by subdisciplines around (Numerical) Mathematics and Information Technologies, which is connected by single links with the first and the third cluster.

Bridging sub-disciplines connecting these four clusters are “Other Natural Sciences”, which link the interdisciplinary Medical Science cluster and the SSH cluster, and “Computer Simulation” together with “Biophysics” connecting Medical research (first cluster), Quantum research (third cluster) and Mathematics (fourth cluster).

Ad 2. To measure the extent of interdisciplinarity in the publications, we applied a special Hill-type indicator ${}^2D^2$ according to Leinster and Cobbold (2012), which takes real values ≥ 1. We furthermore used the 74 subfields according to the revised WoS-based Leuven-Budapest classification scheme (cf. Glänzel et al., 2016) for optimum granularity and compatibility with the above indicator analysis (Glänzel & Thijs, 2018). According to Zhang et al. (2016), values above 5.0 reflect strong interdisciplinarity. Figure 4 gives a detailed comparison between interdisciplinarity scores within the SFB programme and the Austrian standard displayed by major ECOOM fields. While the scores for Physics, Chemistry, Mathematics and Engineering are by and large in line with Austria’s reference scores, the scores for the output of SFB in the life sciences is less stable. The below standard interdisciplinarity in Biology, Biosciences is contrasted by interdisciplinarity scores above the national standard for the smaller fields Neurology and External Medicine (Clin2).



Source: Dinges et al. (2020) based on Web of Science Core Collection

Figure 4: Interdisciplinarity scores for SFB and Austria by science fields

The field comparison provides an explanation for the lower overall SFB scores (2004–2007: 1.88; 2008–2012: 2.01; 2013–2017: 1.98) compared to Austria (2004–2007: 1.99; 2008–2012:

2.08; 2013–2017: 2.09) and the world (2004–2007: 1.94; 2008–2012: 2.05; 2013–2017: 2.08). The publication profile of SFB is in fact more oriented towards Physics, which has overall a lower score than the other fields.

Ad. 3 SFB participants' own assessment of interdisciplinarity is in contrast to the findings of the bibliometric and network analysis. On the one hand, interdisciplinarity is seen as one of the key attributes that makes the SFB attractive and unique in the Austrian R&I funding system for basic research, while on the other hand interviewed SFB participants stress that interdisciplinarity is merely a means to an end, i.e., collaboration between disciplines forms only when and if there is a need for specific expertise, often in the context of using another discipline's analytical tools as opposed to interdisciplinarity as the integration of disciplines.

Survey and interview results highlight that researchers perceive their projects as highly interdisciplinary, which the network analysis and the bibliometric analysis highlight do not fully corroborate. The evidence on this apparent contradiction in opinions on interdisciplinarity shows that researchers.

- Have different understandings of the differentiation between multi- and interdisciplinarity, and a lack of intrinsic motivation for interdisciplinarity;
- Perceive the peer review process as detrimental to highly interdisciplinary projects with high levels of interdisciplinarity among non-closely related fields of science;
- Incorporate strategic considerations of disciplinary distribution in SFB proposals to maximize the chances of success.

Interdisciplinarity is one of the key attributes that makes the SFB attractive. It is understood not only as a strategic aim of the programme, but researchers highlight that the possibility to propose interdisciplinary research topics as among the key motives for SFB participation. The programme's orientation explicitly promoting interdisciplinarity is seen as a unique feature in the Austrian R&I funding system.

Discussion and Conclusions

The evaluation of the Austrian Science Fund's SFB Programme followed a theory-based evaluation approach. Impact pathways were delineated to derive key hypothesis about the contribution of the intervention towards its objectives. A mixed-method approach was pursued towards the empirical analysis to combine different perspectives and increase research integrity. The evaluation resulted in a differentiated picture of the factors contributing to the objective of funding outstanding research.

The analyses performed clearly show that SFB succeeded in supporting exceptional research as evidenced by the outstanding publication and citation record of the funded projects. The publication profile of the SFB programme and its projects in the period 2004–2017 displays a strong continuity in the growth and development of the scientific activity. Furthermore, the publication analysis shows that SFB researchers receive a high citation impact and that their work reflects a very high scientific standard. SFB projects exceed national averages and outperform Stand-alone projects along all metrics of citation impact. Moreover, the bibliometric analysis reaffirms the SFBs' highly successful publication strategy: The PIs of SFB projects in general publish in high impact journals (with respect to their field) and receive more citations than expected for these journals.

The excellence in scientific outcomes is not paralleled by achievements in terms of interdisciplinarity and international collaboration. SFB projects involve fewer internationally co-authored publications compared to the Austrian average. Interdisciplinarity is typically a result of very closely related (sub-) disciplines working in a sub-project. It is mainly found in fields of science, with clearly overlapping content such as biology and medical theoretical sciences and clinical medicine. In terms of average interdisciplinarity scores, the SFB programme lies below the average for Stand-alone projects and the Austrian average.

Acknowledgments

This work was supported by the Austrian Science Fund (FWF) through the project *Evaluation FWF Special Research Programmes (SFB)*. The authors want to thank all members of the evaluation team for valuable comments and discussions on the subject.

References

Amabile, T. M., Patterson, C., Mueller, J., Wojcik, T., Odomirok, P. W., Marsh, M., & Kramer, S. J. (2001), Academic-practitioner collaboration in management research: A case of cross-profession collaboration. *The Academy of Management Journal*, 44(2), 418–431.

Blamey, A., Mackenzie, M. (2016). Theories of Change and Realistic Evaluation: Peas in a Pod or Apples and Oranges? Evaluation. https://doi.org/10.1177/1356389007082129

Carvalho, S., & White, H. (2004). Theory-based evaluation: The case of social funds. American Journal of Evaluation, 25(2), 141–160.

Chen, H-T (1990) Theory-Driven Evaluations. Newbury, CA: SAGE.

Dinges, M., Heller-Schuh, B., Kalcik, R., Scherngell, T., Wang, A., Glänzel, W., & Thijs, B. (2020), *Evaluation FWF Special Research Programmes (SFB)*. Zenodo. http://doi.org/10.5281/zenodo.3889307

Glänzel, W., Thijs, B., Debackere, K. (2019), *Citation classes: a distribution-based approach to profiling citation impact for evaluative purposes*. In: W. Glänzel, H. Moed, U. Schmoch, M. Thelwall (Eds.), Springer Handbook of Science and Technology Indicators. Springer International Publishing - Berlin, Heidelberg, 335–360.

Glänzel, W., Thijs, B. & Chi, P.S. (2016), The challenges to expand bibliometric studies from periodical literature to monographic literature with a new data source: The Book Citation Index. *Scientometrics*, 109(3), 2165–2179.

Glänzel, W., & Thijs, B. (2018), The role of baseline granularity for benchmarking citation impact. The case of CSS profiles. *Scientometrics*, 116(1), 521–536.

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2009), Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, 78(1), 165–188.

Katz, J.S. & Martin, B.R., (1997), What is research collaboration? *Research Policy*, 26, 1–18.

Leinster, T. & Cobbold, C.A. (2012), Measuring diversity: the importance of species similarity. *Ecology*, 93(3), 477–489.

Mayne, J. (2001). Addressing attribution through contribution analysis: Using performance measures sensibly. Canadian Journal of Program Evaluation, 16(1), 1–24.

Mayne, J. (2008). Contribution analysis: An approach to exploring cause and effect.

Mayne, J. (2011). Contribution analysis: Addressing cause and effect. Evaluating the Complex, 53–96.

Powell, W. W. & Grodal, S. (2005), Network of Innovators. In: Fagerberg, J., Mowery, D. C., & Nelson, R. R. (Eds.), *The Oxford Handbook of Innovation.* Oxford: Oxford University Press, 56–85.

Scherngell, T. (Ed.) 2013, *The Geography of Networks and R&D Collaborations. Advances in Spatial Science*. Berlin, Heidelberg: Springer.

Zhang, L., Rousseau, R. & Glänzel, W. (2016), Diversity of references as an indicator for interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the American Society for Information Science and Technology*, 67(5), 1257–1265.

Co-link analysis as a monitoring tool: A webometric use case to map the web relationships of research projects

Jonathan Dudek[1], David G. Pina[2] and Rodrigo Costas[3]

[1] *j.dudek@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (The Netherlands)
Delft Institute of Applied Mathematics, Delft University of Technology, Delft (The Netherlands)

[2] *David.Pina@ec.europa.eu*
European Research Executive Agency, European Commission, Brussels (Belgium)

[3] *rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (The Netherlands)
Centre for Research on Evaluation, Science and Technology (CREST), Stellenbosch University (South Africa)

Abstract

This study explores the societal embeddedness of the websites of research projects. It combines two aims: characterizing research projects based on their weblink relationships, and discovering external societal actors that relate to the projects via weblinks. The study was based on a set of 121 EU-funded research projects and their websites. Domains referring to the websites of the research projects were collected and used in visualizations of co-link relationships. These analyses revealed clusters of topical similarity among the research projects as well as among referring entities. Furthermore, a first step into unveiling potentially relevant stakeholders around research projects was made. Weblink analysis is discussed as an insightful tool for monitoring the internal and external linkages of research projects, representing a relevant application of webometric methods.

Introduction

Publicly funded research projects are a prominent means to develop scientific research. In the context of Horizon 2020 (H2020), the EU's Framework Programme for research and innovation from 2014 to 2020, a broad variety of projects have been funded. Given the scope, size, and duration of many of those, evaluating their impact in a timely manner has to go beyond common metrics for scientific output, such as publications. Those, as well as citations, provide relatively belated signals of the – academic – impact of projects, and usually at a time when projects have already been finished. A more general question is whether funded research projects are raising any interest or attention in society, and whether this can be captured timelier. One of the measures to explore a project's embeddedness in society can be the analysis of the *digital response* that a project's web presence generates. In this, a research project's website can serve as a point of reference to which other entities on the Internet *respond digitally* in the form of weblinks.

Weblinks and webometrics

The field of webometrics has produced a broad body of research studying weblinks, how they form networks, and, among a variety of other questions, how websites can be ranked and measured based on their connections (Thelwall, 2012). However, researchers have also advised caution regarding inferences based on one-sided quantitative analyses of weblinks (Thelwall, 2006). The challenge of adequately characterizing and interpreting weblinks is considerable. While weblinks can be seen as *web-'citations'*, in a way comparable to scientific citations, such equivalence can be problematic when not considering the different motivations and contexts behind weblinks (Björneborn & Ingwersen, 2004). Nonetheless, it has also been argued that weblinks may be useful for covering the phases of research that precede the formal publication of research output (Thelwall, Klitkou, Verbeek, Stuart, & Vincent, 2010).

Using weblinks to characterize the digital response to research project websites

Considering the conceptual caveats around weblinks, it is obvious that the incoming links attracted by a website need to be turned into processable and meaningful metrics. This can be achieved by a combination of measures and may include a focus on the origins of web references (Thelwall, 2006). Furthermore, weblinks can – mimicking citation relationships – be combined in co-citations and bibliographic couplings (Björneborn & Ingwersen, 2004). This might as well be described as "heterogenous couplings" (Costas, de Rijcke, & Marres, 2020) more broadly. As far as heterogenous couplings assume networked structures underlying science-non-science interactions, they may be used for creating networks of web relationships as well. This would account for the fact that websites are (usually) embedded in networks of other web entities and thus, provide an opportunity for contextualizing the unique *digital footprint* of e.g., a project website. At the same time, considering a 'community' of web pages can also reveal insights regarding how much it refers to itself, as well as the extent to which it is being referenced from outside.

Our study aims to apply the concept of heterogenous couplings on the characterization of the weblink relationships of research projects. We investigate the relationships between the websites of research projects based on weblinks among them, as well as weblinks coming from outside their community. Accordingly, this research in progress revisits existing webometric methods and applies them to the analysis of the websites of publicly funded research projects.

Methodology

Our analysis was based on 121 research projects within the *Science with and for Society* (SwafS) program, funded under H2020, representing 77% of all SwafS projects funded by the end of 2020. We selected this portfolio of projects as study case because their objectives are linked to aspects such as science communication, careers and education, promoting gender equality, responsible R&I, and research ethics and integrity, and the engagement of citizens. The type of outputs and outcomes delivered by these projects requires the search or definition of valid forms of assessing their impact and monitoring their activities beyond the typical research related metrics. All projects included had started between 2015 and 2020. At the time of this analysis, 46 (38%) of all projects had ended already, the rest was still running. Each of the projects had a dedicated project website.

Identifying referring sources

To identify the web sites linking to a research project's website(s), we relied upon *backlinks*. Backlinks are the links that a website receives (Björneborn & Ingwersen, 2004). Multiple commercial services crawl the Internet for such backlinks and extract them. We used Majestic, an exhaustive source for webometric research (Orduña-Malea, 2021). This service operates a large database of weblinks and is broadly used for the task of optimizing web presences.[1]

Getting the backlinks for the projects' websites from Majestic, we did not use the path (or "URL") of the homepages of the project websites, since this would only return the backlinks referring to this specific page. Instead, we looked up the *root domain* in a given URL. This expanded the set of results by returning any link received by any webpage available under the root domain.[2] For a given website this means that both backlinks to the homepage and any sub-page are found, providing a more complete picture of the embeddedness of a website. We did not collect all the backlinks, but only focused on the referring domains. This was in order to prevent results from being biased by a large number of backlinks coming from one single

[1] https://majestic.com/company/about

[2] For example, in the URL https://europa.eu/, the part 'europa.eu' is the root domain.

domain (i.e., any link relationship between a project website and a referring website was counted just once). We collected all referring domains accordingly on January 13, 2021.
Since the SwafS projects are funded and conducted in a European context, we limited the referring domains to those originating from an EU member state or an associated country, based on the country codes coming with each referring domain or where the so-called top-level domain (TLD) indicates such a country.[3] This resulted in a final set of 4,606 referring domains. The projects in our set might have had different chances of gaining online attention, depending on the time they had already been running. We tested for a correlation between the number of days a project had been existing from its starting day until January 13, 2021, and the number of referring domains, returning a positive, moderate correlation ($r_s = .491$).

Creating networks based on co-link relationships

To analyze how the SwafS project websites relate to each other, we considered the overlap between referring domains and the project websites' root domains. Per project, we calculated the share of referring domains coming from other SwafS projects among all referring domains. Then, we created a matrix of projects based on the number of other projects that refer to both of them, establishing a *co-linked* relationship between them (see the graph on the left in Figure 1). In the field of bibliometrics, this is also known as a *co-citation* (Björneborn & Ingwersen, 2004; Costas et al., 2020). Next, we created the same matrix of co-linked projects, but this time using the number of *external* domains linking to more than one SwafS project. Finally, we connected the referring, external domains based on mutually linked project websites (see the graph on the right in Figure 1). This *co-linking* is also referred to as *bibliographic coupling* (Björneborn & Ingwersen, 2004; Costas et al., 2020).

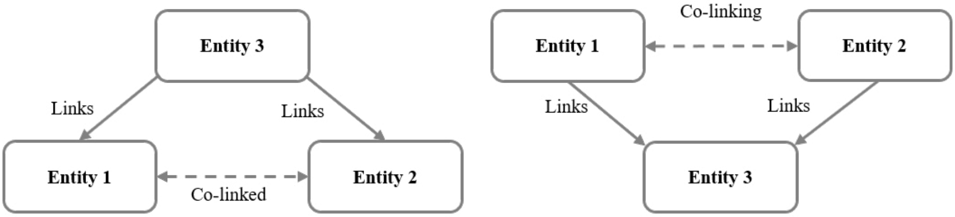


Figure 1. Schemes for a co-linked relation (left), and a co-linking relation (right).

In the networks with external referring domains, we did not include those with a '.com'-TLD. With 805 distinct cases found, this was the most common TLD, followed by '.eu' (424) and '.org' (390). Our reasoning was that domains with this dominant TLD could divert the focus from more interesting actors, which we assumed to rather be present with organization ('.eu', '.org') or country ('nl', 'de', etc.) related domains. For example, 90.1% of the project domains itself have an '.eu'-TLD. All networks were visualized with the VOSviewer software.[4]

Results

Overall, we observed a considerable weblink interconnection within the SwafS projects community. On average, 10.6% of the domains that refer to a SwafS project's website came from other SwafS projects in our set. However, the extent to which this was the case varied a lot: 24 projects were not referred to by any other project, whereas in the case of one project, 43.5% of all referring domains originated from other SwafS projects' websites. The extent of

[3] The top-level domain is the URL part following the so-called second-level domain. E.g., in the URL https://europa.eu/, 'eu' is the top-level domain, and 'europa' is the second-level domain.
[4] https://www.vosviewer.com/

the overlap becomes also apparent when mapping the weblinks among projects based on the number of other projects that refer to them (see Figure 2). What is relevant here are different clusters of projects that are formed based on weblinks. For example, the green cluster on the right refers to projects that are aimed at (gender) equality. The blue cluster (bottom left) can be related to citizen science. In between these clusters, the red and yellow clusters include projects focused on Responsible Research and Innovation (RRI).

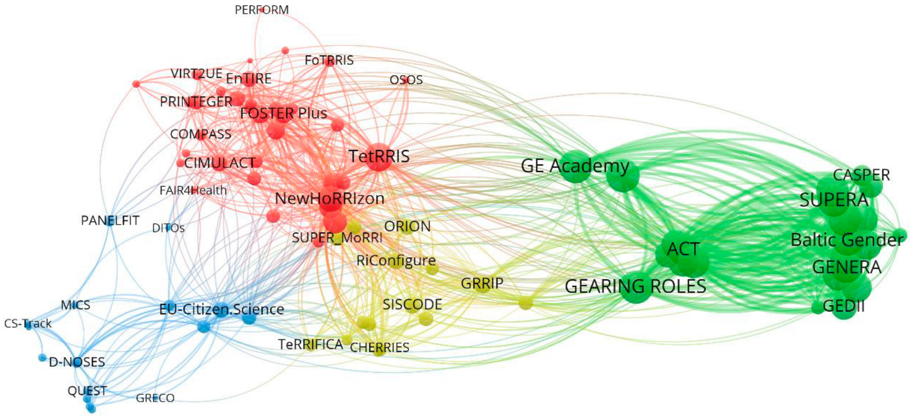


Figure 2. VOSviewer visualization of a co-linked network of research projects based on referring other research projects.

A similar pattern of clusters becomes visible when again looking at the co-linked network for the research projects, but this time only based on external sources of reference (Figure 3).

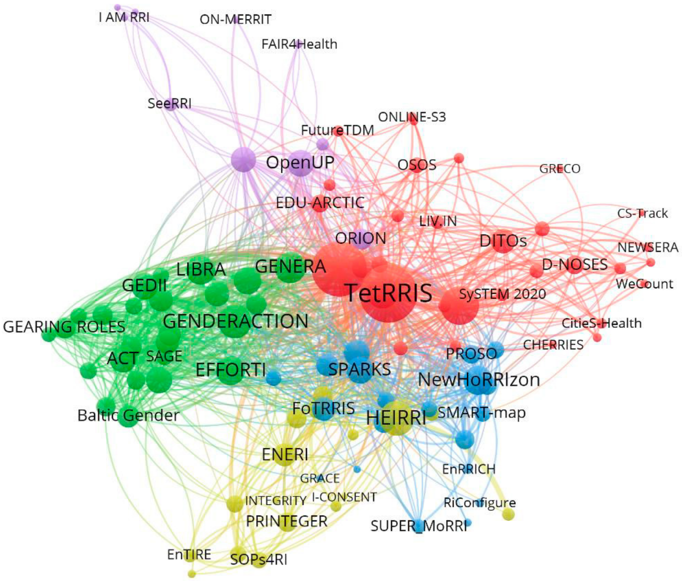


Figure 3. VOSviewer visualization of a co-linked network of research projects based on external referring domains.

Finally, coupling external referring domains based on concurrently cited research projects (Figure 4) revealed several clusters that reflect some of the topics also present in the cluster networks in Figures 2 and 3: For example, we again find a cluster related to gender equality (yellow), and a cluster relatable to citizen science (dark blue). The linking organizations include, among others, universities, science networks, museums, research institutes, and funding organizations.

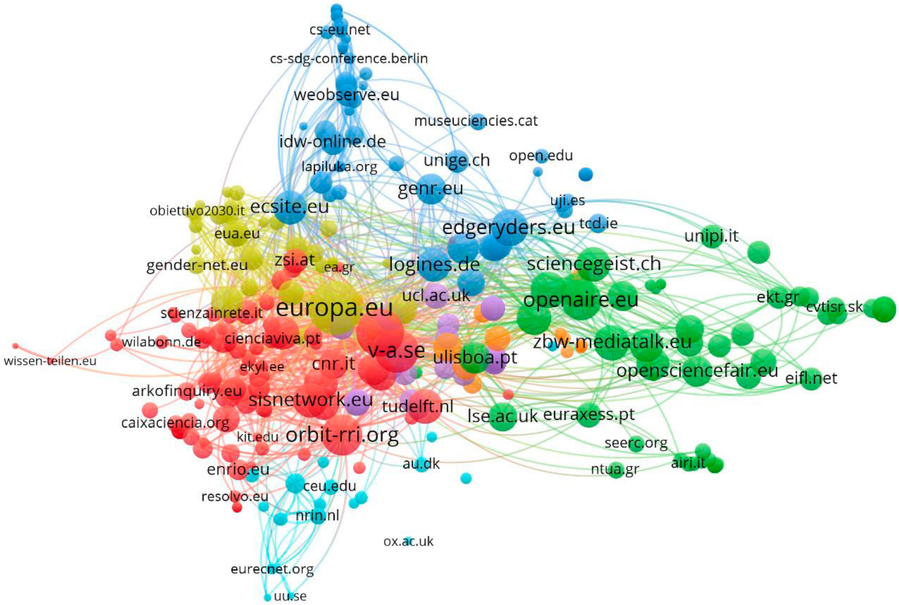


Figure 4. VOSviewer visualization of the co-linking network of referring domains.

Discussion

Our study shows how the weblink relationships of SwafS research projects capture underlying topical clusters. This could provide valuable insights into the relatedness of research projects and hint to possible synergies between them. Our proof of concept also shows the network of external actors based on the weblink connections with the projects. These external linking actors represent a unique source of relevant information on potential societal stakeholders (e.g., governments, museums, networks, etc.) who exhibit an interest in the projects. This hints to the potential value of weblink analyses to assess the ‘early’ societal relevance of (funded) projects, making this also very relevant for funding organizations. They could identify latent relationships existing among their funded projects at an early stage of the projects.[5]

The approach presented has relevance for both the ‘community’ analysis of research projects (i.e., the analysis of a set of projects from the same program or topic), and the analysis of individual projects. Combined with background information on the projects, a weblink analysis could illuminate underlying characteristics of individual projects (e.g., the relationships with other projects, or with other organizations), providing a unique tool to identify potential stakeholders relevant to each project.

[5] Project websites are typically created at the early stages of projects, while other outputs that could also point to project relationships – like reports or publications – generally take more time to be produced.

We conclude that the weblink analysis of scientific projects represents a strong monitoring tool that is able to characterize the internal and external linkages resulting from funded projects. This adds to the variety of applications of webometrics as well, and, more specifically, to existing research on science-related web presences (Orduña-Malea, 2021).

It should be noted that the approach presented is only a *snapshot in time*. The Internet is dynamic, and the numbers of referring domains can change quickly over time. Furthermore, research projects may have different requirements and ambitions regarding their online presence, which is connected to the numbers of links they may possibly receive. This requires caution especially when making inferences for individual cases.

We plan to further refine the approach described, including the analysis and classification of the referring sources to better grasp the underlying structures. In addition to that, insights could be obtained by breaking down results according to the countries of origin of referring sources and by connecting them to the origins of project participants.

We are also going to investigate how the approach described plays out in a different context – namely, on Twitter. For that, we aim to study connections between the Twitter accounts of the research projects and other Twitter accounts. This could eventually lead us to uncovering the "communities of attention" (Haustein, Bowman, & Costas, 2015) around research projects and thus, result in a more complete picture of the societal linkages to research projects.

Acknowledgements

This research is partially funded by the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP). We would like to thank Enrique Orduña-Malea for helpful comments on the draft of this research in progress paper.

Disclaimer

All views expressed in this article are strictly those of the authors and may in no circumstances be regarded as an official position of the European Research Executive Agency or the European Commission.

References

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology, 55*(14), 1216–1227.

Costas, R., de Rijcke, S., & Marres, N. (2020). "Heterogeneous couplings": Operationalizing network perspectives to study science-society interactions through social media metrics. *Journal of the Association for Information Science and Technology, 72*(5), 595–610.

Haustein, S., Bowman T. D., & Costas, R. (2015, July 27-29). *Communities of attention around journal papers: who is tweeting about scientific publications?* [Conference presentation]. Social Media and Society 2015 International Conference, Toronto. https://es.slideshare.net/StefanieHaustein/communities-of-attention-around-journal-papers-who-is-tweeting-about-scientific-publications

Orduña-Malea, E. (2021). Dot-science top level domain: Academic websites or dumpsites? *Scientometrics, 126*(4), 3565–3591.

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology, 57*(1), 60–68.

Thelwall, M. (2012). A history of webometrics. *Bulletin of the American Society for Information Science and Technology, 38*(6), 18–23.

Thelwall, M., Klitkou, A., Verbeek, A., Stuart, D., & Vincent, C. (2010). Policy-relevant Webometrics for individual scientific fields. *Journal of the American Society for Information Science and Technology, 61*(7), 1464–1475.

Early insights into the Arabic Citation Index

Jamal El-Ouahi[1,2]

[1] *j.el.ouahi@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, Netherlands
[2] Clarivate Analytics, Dubai Internet City, Dubai, United Arab Emirates

Abstract

The Arabic Citation Index (ARCI) was launched in 2020. This study gives an overview of the scientific literature available in this new database. I analyse ARCI by using metadata available in scientific publications to characterize its coverage. First, I describe the data and the methods used in the analyses. As of October 2020, ARCI indexed 65,208 records covering the 2015-2019 period. Second, I explore the literature distributions at various levels (research domains, countries, languages, open access). Close to 99% of documents indexed are articles. Results reveal the concentration of publications in the Arts & Humanities and Social Sciences fields. Most journals indexed in ARCI are currently published from Egypt, Algeria, Iraq, Jordan and Saudi Arabia. Around 7% of publications in ARCI are published in languages other than Arabic. Then, I use an unsupervised machine learning model, LDA (Latent Dirichlet Allocation) and the text mining algorithms of VOSviewer to uncover the main topics in ARCI. These methods are particularly useful to better understand the topical structure of ARCI. Finally, I suggest few research opportunities after discussing the results of this study.

Introduction

Research excellence is often equated to publishing in English in high impact factors journals, as stated in the third principle of the *Leiden Manifesto* (Hicks et al., 2015). This is problematic for the humanities and social sciences where research tends to be more engaged on national issues and published in local languages. Identifying the peer-reviewed journals of regional relevance and importance is a major issue for all scientific stakeholders. Protecting excellence in locally relevant scientific research is key to preserve fields which have regional or national dimensions.

In 2020, the Arabic Citation Index (ARCI) was launched in the Web of Science platform, first in Egypt and later in the rest of the Arab World. Clarivate Analytics partnered with the Egyptian Knowledge Bank (EKB), as part of the Egyptian Ministry of Education, to power the first Arabic Citation Index. This launch is also part of Egypt's *Vision 2030* where knowledge, innovation and scientific research are key pillars to achieve scientific excellence (Egyptian Government, 2016). As mentioned by Dr. Shawki, Minister of Education & Technical Education in Egypt and President of the EKB Project, the "aim is to work toward becoming a more knowledgeable Egyptian community that encourages learning as a part of everyday life. We look forward to building our economy and exporting our sciences globally in the Arabic language." (Clarivate Analytics, May 2018).

The focus of ARCI is on the 22 countries of the Arab League and their scholarly research. ARCI joins other regional citation indexes, using the same core features with an Arabic language interface. The criteria for inclusion in ARCI are a subset of the Web of Science Core Collection criteria (Clarivate Analytics, 2019). The journals covered in ARCI are selected by a newly established Editorial board with members from Arab League countries who provide subject knowledge and regional insights. The selection process for ARCI is based on traditional scientific publishing standards and the research norms of the Arab region. First, there is an initial triage to confirm content accessibility and format for all titles considered for indexation in ARCI. All journals must have an ISSN. Several elements are evaluated in this first step: journal title, publisher information, URL for online journals, content access, DOI/pagination and timeliness/volume. Next, the journals are reviewed from an editorial perspective. In this second step, each journal is evaluated to confirm it provides scholarly content, with a clear scope statement, article abstracts, cited references, content relevance with the stated scope or mission, quality of language consistent with scientific communications and an editorial board reflective of the field of the journal.

Such indexing will provide more visibility to research published in Arabic facilitating contribution to local and international research efforts. ARCI is relatively new, and it requires a separate subscription to be accessed in the Web of Science platform. Very little is known about this new database which is still unfamiliar to many researchers. Therefore, the purpose of this study is to characterize the literature available in this new citation index. First, I describe the data and methods used in the analyses. Next, I explore few content distributions at various levels (research domains, countries, languages, open access). Then, I analyse the main topics covered in ARCI by using the Latent Dirichlet Allocation model and the text mining algorithms of VOSviewer (van Eck & Waltman, 2010). Finally, I discuss the results of this study, identify its limitations and suggest few research directions.

Data and Methods

Data

ARCI has a coverage back to 2015. ARCI data was extracted on October 5th, 2020. I have excluded 2020 since the year was not complete yet. Full records were exported from the Web of Science platform. The dataset under study consists of 65,208 records for the 2015-2019 period. Table 1 lists the distribution of records by publication year as well as the share they represent within ARCI.

Table 1. Number and share of ARCI records by year (2015-2019)

Publication Years	*Records*	*Share (%)*
2019	14,198	21.8
2018	15,307	23.5
2017	14,774	22.7
2016	12,013	18.4
2015	8,916	13.7

This database is well structured with 48 fields of information in each record allowing multiple bibliometric analyses (e.g. Publisher Information, Funding Information, Research Area, Open Access Indicator, Cited References, Citations, Usage Counts, ESI Highly Cited Paper/Hot Paper). In addition to essential metadata available in English as in the Web of Science Core Collection, ARCI has some specific information written in Arabic such as *authors names*, *article title*, *publication name*, *author keywords*, *abstract* and *author address.* ARCI records also show the ARCI times cited and the Total Times Cited Count (Web of Science Core Collection, Arabic Citation Index, BIOSIS Citation Index, Chinese Science Citation Database, Data Citation Index, Russian Science Citation Index, SciELO Citation Index) as well as the Cited References and the Cited Reference Count.

Methods

In this study, I use bibliometric methods to characterize the literature indexed in ARCI. The objective is to examine the local research landscape from various perspectives. Such analyses can help research managers and policy makers to better understand local research activity. I conducted a bibliometric analysis to study the journal distribution across countries and scientific productivity of countries. Then, I explored the distribution of publications by research fields, languages and access types.

Many machine learning algorithms have also been developed to understand, group or search information from large text databases. Such approaches have been frequently used to examine the structure of an aggregated literature. In natural language processing, a topic model is a statistical model to discover the abstract hidden semantic structures or topics that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is a model proposed by Blei et al. (2003) used to classify text in a document to a topic. Previous studies have shown that LDA performed well to understand the topical structure of a scientific corpus (Han, 2020; Suominen & Toivanen, 2016; Yau et al., 2014).

LDA is a generative probabilistic model of a corpus. The basic idea is that publications are composed of groups of words with no sequential relationship between them. As documents can include multiple topics, there is a probability of topic distribution for each topic. Each record can be described by a distribution of topics. Each topic is characterized by a distribution over words, described as a distribution of terms in a fixed vocabulary.
There are many tools available for LDA. In this study, I applied the *LatentDirichletAllocation* model available in the Scikit-learn.org python library to perform LDA.

Results

In this section, I report the main findings of the study. First, I analyse the research domains in ARCI by number of records and the proportion they represent in the database. Next, I present the journals distribution by countries. Then, the most productive countries are examined, followed by an analysis of the languages of publications and their access types. Finally, I focus on the main topics covered in the Arabic scientific literature indexed in ARCI.

Research areas distribution

Research Areas constitute a subject categorization scheme that is shared by all Web of Science product databases. This is particularly helpful when analysing documents from multiple databases related to the same research areas. All 153 research areas in the Web of Science are grouped into five broad categories: Arts & Humanities, Life Sciences & Biomedicine, Physical Sciences, Social Sciences, and Technology.

I relied on the journal category and not on the topic covered in the individual publications to analyse the disciplinary coverage in ARCI. These categories or areas, which are defined at the journal level, are used as proxies for scientific fields.

The ARCI records relate to 21 research areas in the dataset under study. Currently, 11,070 records (around 17% of ARCI), do not contain data in the *Research Area* field. In Figure 1, I summarize the number of records within each of research area and the share of the database they represent. I have limited to the 16 research areas with a share higher than 0.5%.

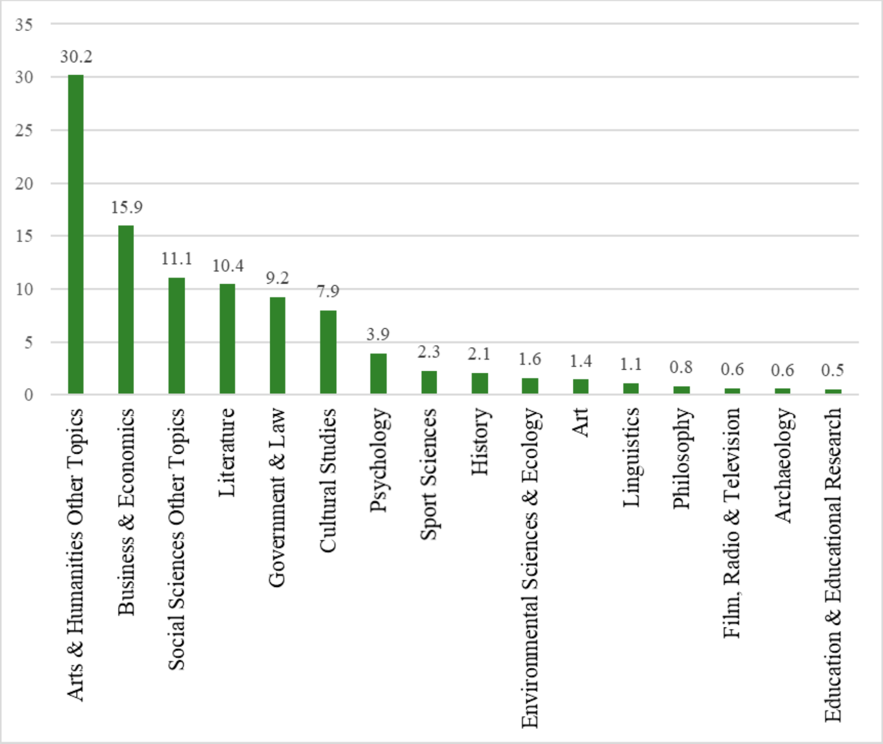


Figure 1. Share (%) of documents by research areas in ARCI (2015-2019)

I have also summarized the shares of the number of papers within each of the five broad domains in Figure 2.

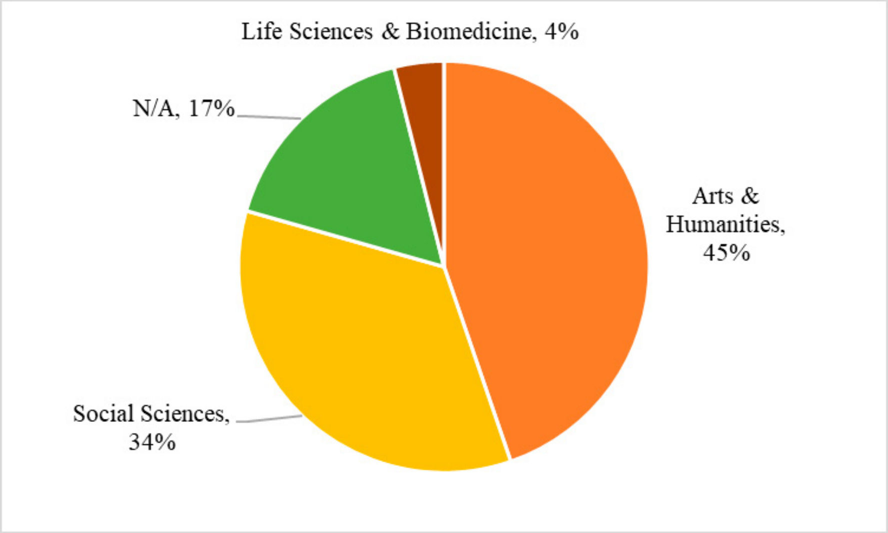


Figure 2. Share (%) of documents by broad categories in ARCI (2015-2019)

This figure shows ARCI contains mainly journals in the Arts & Humanities, and Social Sciences categories. These categories represent 79% of ARCI total coverage. Journals in Life Sciences & Biomedicine account for 4% of the coverage. As mentioned earlier, 17% of records retrieved do not contain information about the Research Area. It is worth noting there is no journal related to Technology or Physical Science categories. This confirms the current focus of ARCI. Local issues in Arts & Humanities as well as Social Sciences dominate the ARCI coverage.

ARCI also offers its own research categories. We retrieve similar results. However, some differences emerge. When analysing the records with the ARCI classification, only 242 records do not contain information about the research categories, representing less than 0.4% of the total database. Some categories standout such as Islamic Studies, Islamic Jurisprudence, Islamic Creed, Poetry and Hadith which are fields well studied in the Arab region.

Content coverage by countries

In this section I analyse the coverage by country. First, I examine the types of documents indexed in ARCI. Table 1 lists the number of documents per type and they share they represent in the database. ARCI is primarily composed of articles. Close to 99% of documents indexed are articles. Other document types all represent less than 1% of the database.

Table 1. Number and share of ARCI records by document type (2015-2019)

Document Type	*Records*	*Share (%)*
Article	64,467	98.864
Review	432	0.662
Editorial	104	0.159
Art and Literature	97	0.149
Other	62	0.095
Bibliography	43	0.066
Meeting	3	0.005

Now, I focus on the distribution of journals over countries, where each journal is assigned to a country based on the country in which the publisher is located. Before analysing the country distribution in ARCI, I examined the coverage of Arab journals in the various citation indices in the Web of Science Core Collection (WoS CC): Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI) and Emerging Sources Citation Index (ESCI). This coverage is represented in Figure 3.

As of October 2020, 21,419 journals were indexed in WoS CC. 144 journals (or 0.67%) are published in 13 of the 22 Arab League countries: 66 in SCIE, 1 in SCIE and SSCI and 77 in ESCI. Out of these 144 journals, 134 (93%) are published in English only. The remaining 10 journals (7%) have published papers in several languages during the study period: English (78%), French (12.6%), Arabic (5.8%), Spanish (3.365%), Afrikaans (0.05%) and Italian (0.05%). The United Arab Emirates (UAE), Egypt and Saudi Arabia are the three most represented Arab countries in WoS CC with a total of 113 journals and a share of 78% of all journals published in the Arab region and indexed in WoS CC.

Although the criteria for inclusion in ARCI are a subset of the Web of Science Core Collection selection process, there is no overlap between WoS CC and ARCI. The distribution by country of publisher in ARCI is represented in Figure 4. ARCI indexes content from 19 of the 22 Arab League countries. Content for Djibouti, Comoros and Somalia is not indexed yet. Journals published in Egypt, Algeria, Iraq, Jordan and Saudi Arabia represent 79% of the journals indexed in ARCI.

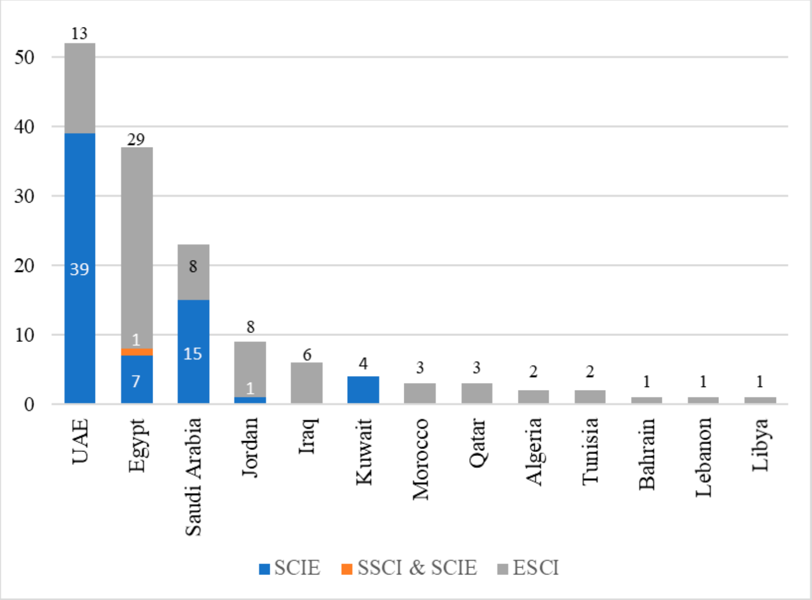


Figure 3. Number of journals by Arab country in WoS CC citation indices (September 2020)

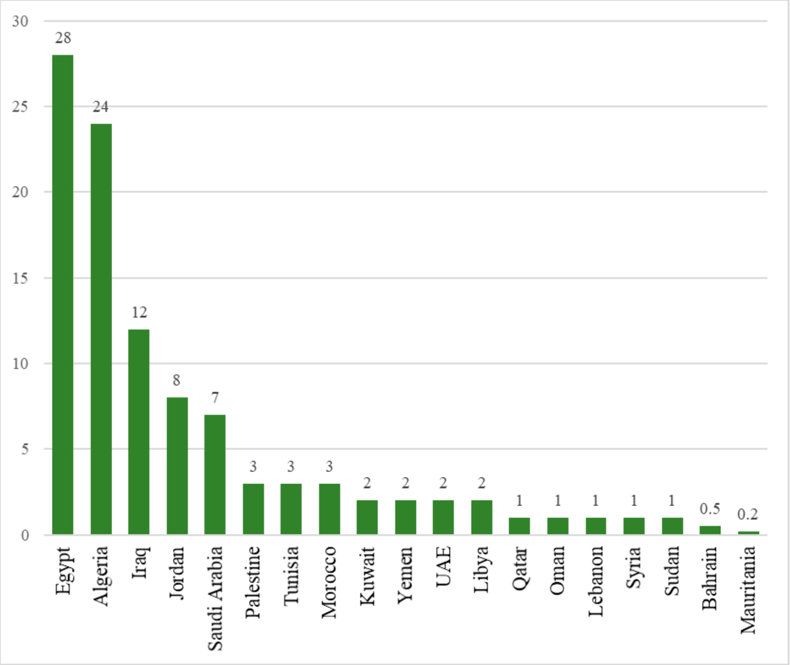


Figure 4. Share (%) of journals by country in ARCI (2015-2019)

Egypt and Algeria cumulate more than half of the journals indexed in ARCI. There is currently a concentration of journals published from these two countries in ARCI. The submission process is managed by the Egyptian Knowledge Bank (www.arcival.ekb.eg) and journals are evaluated according to the ARCI selection process as explained in the introduction of this paper. Several Arab countries have also set up national journal platforms and initiatives. For example, the Algerian Scientific Journal Platform (www.asjp.cerist.dz) and the portal of Moroccan scientific journals (www.revues.imist.ma) have been developed by their respective Ministry of Higher Education and Research. The common goal of such initiatives is to improve the visibility of local journals by improving their publishing standards. ARCI is still new and is still growing. It will be interesting to analyse how this new citation index evolves over time in terms of content coverage by journals' countries.

Languages coverage

Table 2 shows the coverage of records in terms of language of publications in ARCI. Arabic obviously dominates the database with 60,439 publications, representing a share of around 93%. The second most represented language is English with 3,221 records (4.94%) then followed by 1,406 publications in French (2.15%). 142 publications in 9 other languages represent 0.22% of this database.

Table 2. Number and share of records by language in ARCI (2015-2019)

	Language	*Records*	*Share (%)*
1	Arabic	60,439	92.69
2	English	3,221	4.94
3	French	1,406	2.16
4	Spanish	35	0.05
5	German	32	0.05
6	Persian	19	0.03
7	Hebrew	18	0.03
8	Italian	16	0.03
9	Russian	11	0.02
10	Chinese	5	0.01
11	Turkish	5	0.01
12	Amazigh	1	–

As ARCI aims to provide more exposure to journals published in the Arab League countries, it is no surprise to see Arabic as the dominant language. It is also worth reminding many journals indexed in ARCI provide publication in multiple languages. Several countries from the Arab League are former British or French colonies which explains why English and French are the main non-Arabic languages in ARCI. Other languages suggest that the research published in ARCI journals might also tackle regional issues of interest with neighbour countries.

Open access

The last few years have seen the development of several open access (OA) options (Bosman & Kramer, 2018; Lewis, 2012). Several scholars have studied the advantage of OA in terms of readership as well as citation impact (Basson et al., 2020; Cintra et al., 2018; Morillo, 2020; Piwowar et al., 2018; Riera & Aibar, 2013; Tang et al., 2017; Torres-Salinas et al., 2019; Young & Brandes, 2020). Since 2014, Web of Science has provided information to identify publications from OA journals. I use this information to analyse the access type of records indexed in ARCI. The statistics for various OA types and non-OA records in ARCI are presented in Table 3.

Close to 24% of papers indexed in ARCI and published between 2015 and 2019 are openly accessible. This is below the share of 31.8% of OA documents in the Web of Science Core Collection for the same period. We notice the various OA types have different shares in ARCI. *Bronze* is the main OA type with 11,029 papers representing close to 17% of ARCI. *DOAJ Gold* has the second highest OA share (4.49%) in this database with 2,927 papers published with this OA type.

Table 3. Number and share of records by Open-Access type in ARCI (2015-2019)

Open access type	*Records*	*Share (%)*
Non-OA	49,405	75.77
All Open Access	*15,772*	*24.19*
Bronze	11,029	16.91
DOAJ Gold	2,927	4.49
Other Gold	1,190	1.82
Green Published	623	0.96
Green Accepted	3	0.00
Unknown (no DOI)	31	0.00

Main topics

Before applying LDA, one must define the number of topics for the corpus. One option is to examine the performance of text clustering on a small dataset. Another way is to choose the number of topics based on judgments or tests (Blei et al., 2003). In this study, the corpus is organized into 10 topics which are listed in Table 4.

The model is applied on the combinations of words available in the title, abstract and author keywords of all records indexed in ARCI. I limited my study to words written in Roman script alphabets. All records have at least the title written in English. 84% of the publications found in ARCI have an abstract in English. Titles, abstracts and keywords written in Arabic and other non-Roman script languages are not analysed in this paper.

Table 4. 10 topics found in ARCI (2015-2019)

#	*Topic*	*#*	*Topic*
1	design islamic role	6	students skills learning
2	social media study	7	algeria economic study
3	iraq study administrative	8	physical effect players
4	saudi arabia education	9	comparative study law
5	children psychological relationship	10	international model heritage

It is straightforward to interpret the topics generated by the LDA model. The results are useful to understand the topical structure of ARCI by highlighting the main topics covered in the Arabic scientific literature indexed in ARCI.

Term map

When applying the LDA model on a corpus, it is assumed one document can address multiple topics. As shown in Table 4, this is helpful to have a precise understanding of the topical structure of a large corpus. However, it does not map the relationships between topics. The purpose of building a so-called term map of the publications in ARCI is to further clarify their contents. I used VOSviewer (van Eck & Waltman, 2010) to create such map.

Titles, abstracts as well as author keywords have been concatenated into a single string which has been used by the text mining algorithms of VOSviewer. I have limited this analysis to the terms which occur at least 15 times. 954 keywords have satisfied this threshold. For each of the 954 keywords, the total strength of the co-occurrence links with other keywords has been calculated by VOSviewer. Figure 5 shows the co-occurrence network for all the terms indicating for each pair of terms the number of papers in which these terms appear together.

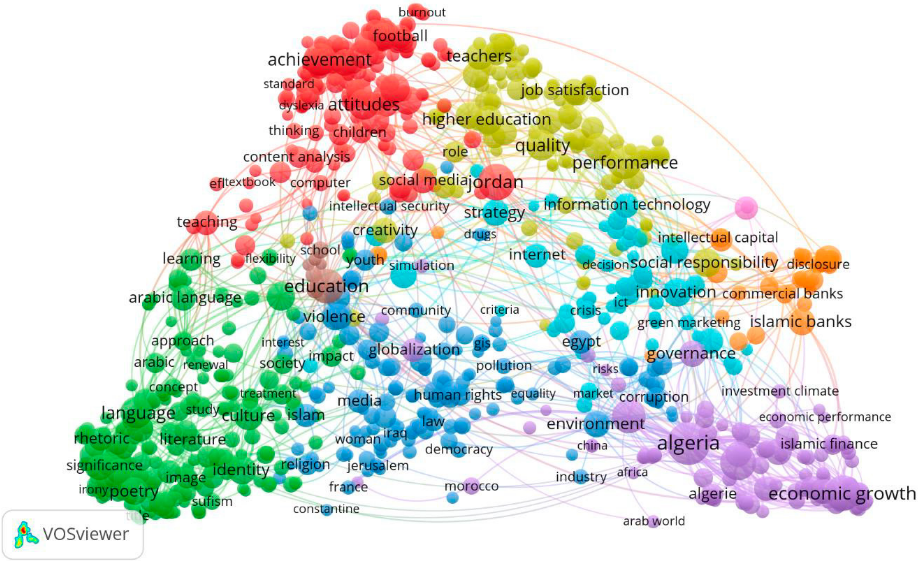


Figure 5. Term map highlighting the main topics in ARCI (2015-2019)

The clustering is useful in delineating the topics covered as well as highlighting the relatedness between them. The horizontal and vertical axes have no meaning. The size of a term reflects the number of records in which this specific term is mentioned. The proximity of two terms is an indicator of how these terms are related based on the number of co-occurrences. In general, groups of terms closely located together can be interpreted as topics. For readability purpose, labels are shown only for selected terms to avoid overlapping labels.

The terms have been clustered into 9 clusters with different colours. The map confirms a broad coverage of scientific literature as shown previously in the topic analysis. The term map shown in Figure 5 indicates some clear distinctions between research areas. These distinctions are not only visible in the structure in terms of proximity between terms but also in terms of colours. Within an area of the map, terms are usually coloured in a consistent way. For example, the lower right part in purple includes research areas related to Economics which are also closely related to the parts in orange (Finance) and teal (Strategy and Innovation). In the upper part of the map, the parts in red, yellow, and brown are more related to Education in general with some distinction on several aspects such as *achievement*, *attitudes, teaching* (in red), *higher education, quality, performance* (in yellow), and *school* and *education* (in brown). The lower left part in green corresponds to research areas related to Literature such as *language, petry, rhetoric, Arabic language, culture* and *identity.* Finally, terms corresponding to the field of Law tend to be located mainly in the blue part of the map with terms such as *law, human rights, democracy, equality, violence* and *community.* There is a clear heterogeneity in terms of topics covered in ARCI. One should remember some of the terms can of course be used in various contexts.

Discussion

The main objective of this study was to examine the structure of ARCI, the first Arabic Citation Index. 290 Arabic journals are indexed in ARCI as of October 2020. This indexation brings several benefits to the scientific community. This new index will improve the visibility of Arabic journals by making them more accessible. All journals indexed in ARCI need to meet selection criteria and essential publications metadata are provided. Such a database could greatly enhance the scholarly literature search. As a result, this will also help researchers to identify critical and influential research published in Arabic by providing access to highly curated peer-reviewed scientific content. Since research evaluation increasingly implies the bibliometric analysis of research output (Wilsdon et al., 2015; Wouters et al., 2015), ARCI could also provide useful bibliometric data sources to research managers for science assessment and research analysis. This would be helpful to *identify and reward excellence in locally relevant research* (Hicks et al., 2015).

Now, I discuss in detail the main findings identified in this analysis. This study reveals that ARCI contains mainly journals in the Arts & Humanities, and Social Sciences categories. These categories represent 79% of ARCI's total coverage. ARCI has a strong focus on local issues in Arts & Humanities as well as Social Sciences. It is important to keep in mind the well documented limitations on subject delineation where I used research areas of journals as proxies of categories to characterize the subject coverage.

As of October 2020, ARCI indexes content from 19 of the 22 Arab League countries, with more than half of the journals indexed in ARCI being published in Egypt and Algeria. Since ARCI is still new and under development, it will be interesting to track its evolution over time. It would also be interesting to analyse the contribution of each Arab League country to ARCI by using the author affiliations. It is worth reminding that the country of publisher of the journal is considered for its indexation in ARCI and not only the language of publication. Thus, ARCI does not include yet the journals published in Arabic in countries not members of the Arab League. As of now, most of the content found in ARCI is composed of articles (98.8% of the database). Since the Humanities tend to traditionally rely on book chapters and books, it will be interesting to analyse the evolution of the coverage by document type. Unsurprisingly, ARCI has a great share of papers published in Arabic (93% of the database). However, English and French are two other languages well represented in ARCI. Other languages suggest research published in ARCI journals may also tackle regional issues of interest with neighbour regions such as Europe and Asia.

Around 24% of the content indexed in ARCI is openly accessible. This share is below the proportion of Open Access (OA) documents in the Web of Science Core Collection for the same period. The OA information available in ARCI is particularly useful to better share scientific knowledge as well as to track the adoption of local OA mandates by research managers. The Global Open Access Portal (GOAP) presented a snapshot of the status of Open Access (OA) to scientific information worldwide. As identified in the GOAP, the Arab States face challenges but also opportunities (Unesco, 2016). Low level of awareness of the OA potential for researchers, publishers and policy makers tops the list of challenges. Lack of policy regulation, research funders' OA mandates and resources to manage OA projects also contribute to the low OA penetration in the Arab world. Nevertheless, several projects and initiatives have been undertaken to promote OA in the Arab region. 70 experts and policy specialists from several Arab Countries met in September 2015 to develop strategies to implement open access to scientific information and research in the Arab countries (Unesco, 2015). The Directory of Free Arab Journals (DFAJ), the first Arab directory of Open Access Journals which provides access to journals published by 172 publishers from 17 Arab countries, is also an example of such initiative.

The topic analysis as well as the term map are helpful to better understand the underlying structure of ARCI. Such techniques provide a great overview of the topics covered in this database. Overall, the clusters found with VOSviewer seem to be closely related and show a broad coverage of ARCI. The title, abstract and keywords in Arabic were not included in the topic analysis. It might be interesting to characterize the literature in ARCI by focusing on the Arabic content as well.

In conclusion, this paper offers a brief profile of the newest citation index in the Web of Science. This paper contributes also to the literature on regional citation indices (Huang et al., 2017; Jin & Wang, 1999; Leydesdorff & Jin, 2005; Moskaleva et al., 2018; Pajic, 2015; Seol & Park, 2008; Velez-Cuartas et al., 2016). One common purpose of such regional databases is to provide more visibility to local journals. ARCI is likely to have positive effects on regional research discovery as well as research management in the Arab region. Future research may seek to propose detailed mappings of ARCI to better understand its structure and impact. Finally, it will also be interesting to track its development and evolution by using dynamic topic models to study the time evolution of topics by using the text available in English as well as Arabic.

Competing Interests

The author is an employee of Clarivate Analytics, the provider of the Web of Science and the Arabic Citation Index.

References

Basson, I., Blanckenberg, J. P., & Prozesky, H. (2020). Do open access journal articles experience a citation advantage? Results and methodological reflections of an application of multiple measures to an analysis by WoS subject areas. *Scientometrics*, 26.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Bosman, J., & Kramer, B. (2018). Open access levels: a quantitative exploration using Web of Science and oaDOI data. PeerJ.

Cintra, P. R., Furnival, A. C., & Milanez, D. H. (2018). The impact of open access citation and social media on leading top Information Science journals. *Investigacion Bibliotecologica, 32*(77), 117-132.

Clarivate Analytics. (2019). Introducing the Arabic Citation Index. Retrieved from https://clarivate.com/webofsciencegroup/campaigns/arabic-citation-index/

Clarivate Analytics. (May 2018). Clarivate Analytics Partners with the Egyptian Knowledge Bank to Power the First Arabic Citation Index. Retrieved from https://www.prnewswire.com/news-releases/clarivate-analytics-partners-with-the-egyptian-knowledge-bank-to-power-the-first-arabic-citation-index-300655525.html

Egyptian Government. (2016). Egypt's Vision 2030. Retrieved from https://mped.gov.eg/EgyptVision?lang=en

Han, X. (2020). Evolution of research topics in LIS between 1996 and 2019: an analysis based on latent Dirichlet allocation topic model. *Scientometrics, 125*(3), 2561-2595.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429-431.

Huang, Y., Zhu, D., Lv, Q., Porter, A. L., Robinson, D. K. R., & Wang, X. (2017). Early insights on the Emerging Sources Citation Index (ESCI): an overlay map-based bibliometric study. *Scientometrics, 111*(3), 2041-2057.

Jin, B., & Wang, B. (1999). Chinese science citation database: Its construction and application. *Scientometrics, 45*(2), 325-332.

Lewis, D. W. (2012). The Inevitability of Open Access. *College & Research Libraries, 73*(5), 493-506.

Leydesdorff, L., & Jin, B. H. (2005). Mapping the Chinese Science Citation Database in terms of aggregated journal-journal citation relations. *Journal of the American Society for Information Science and Technology, 56*(14), 1469-1479.

Morillo, F. (2020). Is open access publication useful for all research fields? Presence of funding, collaboration and impact. *Scientometrics, 125*(1), 689-716.

Moskaleva, O., Pislyakov, V., Sterligov, I., Akoev, M., & Shabanova, S. (2018). Russian Index of Science Citation: Overview and review. *Scientometrics, 116*(1), 449-462.

Pajic, D. (2015). The Serbian Citation Index: Contest and Collapse. In A. A. Salah, Y. Tonta, A. A. A. Salah, C. Sugimoto & U. Al (Eds.), *Proceedings of Issi 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference* (pp. 604-605). Leuven: Int Soc Scientometrics & Informetrics-Issi.

Piwowar, H., Priem, J., Lariviere, V., Alperin, J. P., Matthias, L., Norlander, B., et al. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *Peerj, 6*.

Riera, M., & Aibar, E. (2013). Does open access publishing increase the impact of scientific articles? An empirical study in the field of intensive care medicine. *Medicina Intensiva, 37*(4), 232-240.

Seol, S. S., & Park, J. M. (2008). Knowledge sources of innovation studies in Korea: A citation analysis. *Scientometrics, 75*(1), 3-20.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(10), 2464-2476.

Tang, M., Bever, J. D., & Yu, F.-H. (2017). Open access increases citations of papers in ecology. *Ecosphere, 8*(7).

Torres-Salinas, D., Robinson-García, N., & Moed, H. F. (2019). Disentangling gold open access *Springer handbook of science and technology indicators* (pp. 129-144): Springer.

Unesco. (2015). First Regional Pan-Arab Consultation on Open Access to Scientific Information and Research.

Unesco. (2016). Global Open Access Portal. Retrieved from http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/access-by-region/arab-states/

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538.

Velez-Cuartas, G., Lucio-Arias, D., & Leydesdorff, L. (2016). Regional And Global Science: Publications From Latin America And The Caribbean In The Scielo Citation Index And The Web of Science. *Profesional De La Informacion, 25*(1), 35-46.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., et al. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*.

Wouters, P., Thelwall, M., Kousha, K., Waltman, L., de Rijcke, S., Rushforth, A., et al. (2015). The metric tide. *literature review, Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management, HEFCE, London*.

Yau, C.-K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics, 100*(3), 767-786.

Young, J. S., & Brandes, P. M. (2020). Green and gold open access citation and interdisciplinary advantage: A bibliometric study of two science journals. *Journal of Academic Librarianship, 46*(2).

Network Effects and Research Collaborations: Evidence from IMF Working Paper Co-authorship

Dennis Essers[1], Francesco Grigoli[2] and Evgenia Pugacheva[3]

[1] *dennis.essers@nbb.be*
National Bank of Belgium, Economics and Research Department, Blvd de Berlaimontlaan 14, 1000 Brussels (Belgium)

[2] *fgrigoli@imf.org*
International Monetary Fund, Research Department, 700 19th St NW, Washington, DC 20431 (United States)

[3] *epugacheva@imf.org*
International Monetary Fund, Research Department, 700 19th St NW, Washington, DC 20431 (United States)

Abstract

An important trend in knowledge generation and diffusion is that the co-authorship of research publications has become remarkably more frequent. In this paper we study the role of co-authorship networks for starting and maintaining research collaborations. Relying on the network of the International Monetary Fund (IMF)'s Working Papers--which reflects well the endogenous nature of research collaborations--we document the presence of many authors with few direct co-authors, yet indirectly connected through short co-authorship chains. Two researchers are more likely to team up if their distance is shorter, likely reflecting reduced matching frictions. In addition, productive authors and authors with different co-author network sizes collaborate more, because of synergies between senior and junior researchers. Being employed in the same department and having citizenship of the same region also influence the decision to collaborate. We argue that incentives should be directed to researcher pairs that are initially more distant from each other in the co-authorship network.

Introduction

Knowledge and ideas are widely considered key contributing factors to long-term economic development. Their non-rival nature gives rise to increasing return to scale, which is the underpinning of endogenous growth (Romer, 1990). One major channel through which knowledge and ideas are generated and diffused is the publication of research papers, and an important trend in this respect is that papers are increasingly produced through collaborations across co-authoring researchers (Wuchty, Jones & Uzzi, 2007; Kuld & O'Hagan, 2018). Co-authorship has become a common feature of research publishing because teaming up allows researchers to exploit synergies and economies of scale by pooling ideas, skills, time, and funds (Liu et al., 2020). Multiple co-authorship links, in turn, give rise to collaboration networks, which play a crucial role in both knowledge generation and diffusion. For example, collaborating authors can combine their ideas to innovate, learn from each other, and then pass on the newly acquired knowledge to their current and future co-authors. Indeed, the empirical evidence suggests that knowledge spills over directly to co-authors (Azoulay, Graff Zivin & Wang, 2010; Borjas & Doran, 2015), as well as indirectly to the broader network of collaborating researchers (Hsieh et al., 2018).

This paper sheds light on how such collaboration networks are structured, formed, and maintained. More specifically, we study network properties; investigate the determinants of starting a new research collaboration; and assess whether these are the same factors that lead researchers to continue working together. For these purposes, we construct the co-authorship network of the near-universe of International Monetary Fund (IMF) Working Papers published over almost three decades (1990-2017) and combine it with detailed information on authors' employment history and demographic characteristics.

Several aspects make our dataset an ideal testing ground to study the endogenous nature of how collaborations between researchers are initiated and maintained. First, being one of the main research outlets of the IMF, the Working Paper series is widely read by central bank and government officials, academics, and researchers at think tanks. Working papers therefore contribute to broad-based knowledge diffusion, possibly more than peer-reviewed journal articles (often behind paywalls), which tend to be the focus of the existing literature on co-authorship formation. Second, compared to other IMF research publications, working papers leave more freedom to the authors: they reflect the authors' personal views rather than the institutional ones, can cover the geographical area and topic of interest of the authors, and do not restrain authorship to people within the same division, department, or even the IMF. Third, IMF staff--the main authors of the IMF Working Papers--does not arguably face the same "publish-or-perish" incentives of academia. Hence, we expect collaborations in our dataset to reflect personal preferences (e.g., researchers sharing a similar background) and practical considerations (e.g., researchers working in the same department), more than strategic ones (e.g., co-authorship with internationally renowned researchers that can help get the papers into top-tier journals and advance peoples' careers). Fourth, by focusing on working papers rather than on peer-reviewed journal articles, the problem of long publication lags distorting our findings is much less severe. Finally, we take advantage of our rich demographic and employment information on authors, absent from earlier studies such as Fafchamps, van der Leij & Goyal (2010), to test additional hypotheses about factors influencing research collaboration decisions and to control for possible confounders when making inference on the effect of pure network variables on the probability of collaborating.

The rest of the paper is structured as follows. We first explain the construction of our dataset and present key stylized facts about the co-authorship network of IMF Working Papers. We then describe our empirical strategy to evaluate the determinants of first and subsequent collaborations across authors, and discuss the econometric results. The final section summarizes our main findings.

Data and stylized facts

Dataset construction

We collected information on 6,152 IMF Working Papers written between 1990 and 2017 by 3,918 distinct authors. Based on the information obtained from the IMF website catalog and the text of the publications, for each working paper we compiled the list of authors, the year of publication, and the *Journal of Economic Literature* (JEL) codes used to classify publications by economic research area(s). JEL codes are available starting in 1997.

We complemented the information from the working papers with employee-level data for the years 1999 to 2017. This unique dataset, obtained from the IMF's Human Resource Department, includes annual information on the year of birth (in five-year intervals), tenure, department, citizenship (by region), and gender of IMF staff. We used fuzzy matching and performed extensive manual checks to merge the data.

For non-IMF authors of IMF Working Papers published between 1997 and 2017, we compiled information on their main affiliation based on email addresses where available from the papers, or manual search online in authors' CVs, LinkedIn profiles, and other websites. We organized these external affiliations into six mutually exclusive categories: universities, other international organizations, central banks, government agencies, non-profit organizations, and private companies.

We also collected citation counts for the working papers from Google Scholar. Relative to other sources of citations--such as Citations in Economics (CitEc), Web of Science, or Scopus--Google Scholar has broader coverage of citations in books, dissertations, and in the non-scholarly "grey" literature (e.g., government reports), as well as in non-English language publications (Martín-Martín et al., 2018). The citations we extracted are cumulative as of January 2019. To account for the fact that working papers were released at different times, we annualize the citation count by dividing it by the number of years since publication.

Finally, we define the universe of "active" potential co-authors in any given year t as S_t. This includes all authors since their earliest year of publication, t_0, until their latest year of publication, t_2. A pair of authors $\{i, j\} \in S_t$ is considered active from the year in which both authors are active, $t^{i,j}{}_0 = max\{t^i{}_0, t^j{}_0\}$, until the year in which either one of the two authors publishes his or her last paper, $t^{i,j}{}_2 = min\{t^i{}_2, t^j{}_2\}$. First collaborations are traced with a dummy variable, $fc^{i,j}{}_t$, that is defined only when the author pair is active, and that takes value one when authors i and j collaborate for the first time, and zero for the years before that. Subsequent collaborations are traced with another dummy variable, $sc^{i,j}{}_t$, which is defined only during the period in which the pair is active and after the first publication, and that takes value one in the year or years in which the pair publishes any other paper but the first, and zero otherwise.

Co-Authorship Network

Figure 1 provides an overview of the structure of the IMF Working Paper co-authorship network. To ease the visualization, we focus on the "core" network, which consists of those authors with a minimum of 10 publications and at least 10 different co-authors over 1990–2017. The nodes represent individual authors and the edges joining them represent co-authorship of a publication. The size of the nodes is proportional to the number of publications an author wrote over the entire period, including single-authored papers. The edge width is proportional to the number of repeated collaborations between two authors. Node grey tones refer to authors' affiliations, which are divided into the IMF (light grey) and the six other categories mentioned above (darker shades of grey).

The core author-level network is unsurprisingly dominated by IMF staff. Even in the full network (i.e., including authors with less than 10 publications and/or 10 co-authors), IMF staff accounts for the largest number of unique authors (54.2 percent) and for the largest number of publications by those authors. Academics, which include resident and visiting scholars at the IMF, come a distant second (24.6 percent), followed by central bankers (6.6 percent). Staff from other international organizations (2.8 percent) and authors affiliated with government agencies (1.8 percent), private companies (1.3 percent), or non-profits (0.9 percent) occupy a much less prominent role in the network.

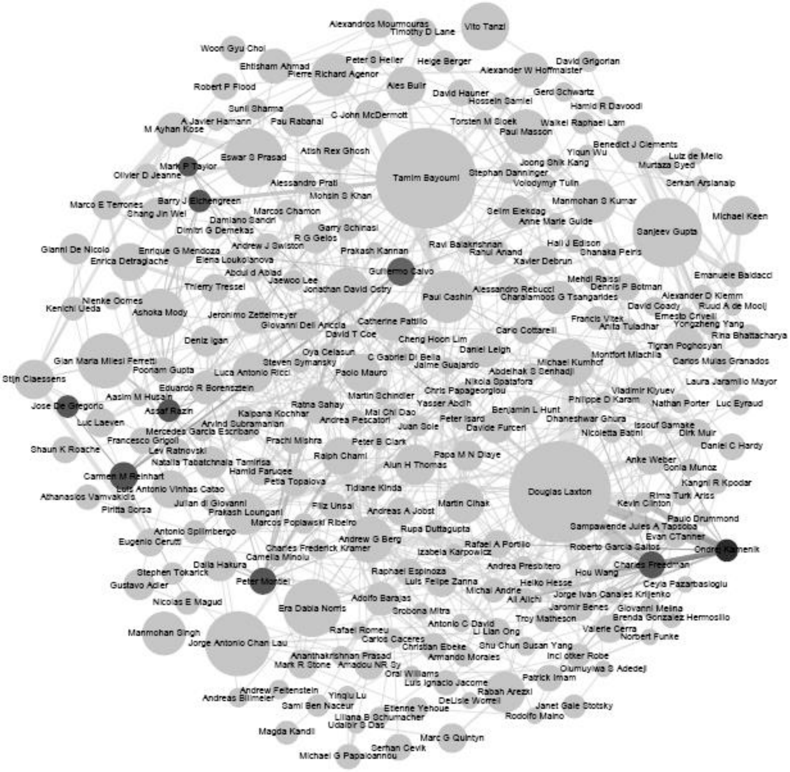


Figure 1. Core author-level network, 1990-2017

To gain further understanding of the network structure, Table 1 reports the key summary statistics generally used to describe networks. As shown in column (1), the full IMF Working Papers network counts 3,918 nodes and 9,027 edges, implying an average degree, or number of co-authors per researcher, of about 4.6. The network has a giant component (i.e., the largest group of nodes that are all directly or indirectly connected to each other) that covers 3,206 nodes, or almost 82 percent of all authors in the entire network. In other words, there are relatively few isolated authors or author groups. Moreover, research at the IMF does not seem to consist of distinct "islands" of researchers. The average distance, or length of the shortest path between two authors in the giant component of the network, is relatively short at about five. The average clustering coefficient, which measures the overlap in co-authorship (defined as the fraction of an author's collaborators that are themselves also co-authors, averaged over all authors with a minimum degree of two) equals more than 0.6.

Table 1. Network statistics

	IMF Working Papers				EconLit
	1990-2017 (1)	1990-1999 (2)	2000-2009 (3)	2010-2017 (4)	1990-1999 (5)
Total nodes (authors)	3,918	923	1,891	2,192	81,217
Total edges (co-authorship)	9,027	1,257	4,081	6,058	67,897
Average degree	4.608	2.724	4.316	5.527	1.672
Size of giant component	3,206	594	1,598	1,916	33,027
Share of giant component (percent)	81.8	64.4	84.5	87.4	40.7
Average distance in giant component	4.961	5.132	4.878	4.640	9.470
Average clustering coefficient	0.614	0.443	0.504	0.613	0.157

Columns (2) to (4) show the same summary statistics for the networks of authors that published IMF Working Papers within each of three sub-periods: 1990-1999, 2000-2009, and 2010-2017. The co-authorship network has grown larger and denser over time. The number of authors more than doubled from just over 900 during 1990-1999 to almost 2,200 during 2010-2017. Meanwhile, the number of co-authorship links nearly quintupled between the first and the third period, resulting in an increase of the average degree from 2.7 to 5.5. During the 1990s, the giant component already covered 64 percent of the authors publishing during that period, but this further increased to almost 85 percent in the 2000s and even more than 87 percent in the 2010s. The giant component also became more integrated, with an average distance progressively declining from 5.1 to 4.6 between the 1990s and 2010s. In addition, the extent of clustering rose considerably. Part of this can be ascribed to the higher share of multi-authored publications in recent years, which (by definition) increases the average clustering coefficient.

To put the attributes of the IMF Working Papers co-authorship network in perspective, column (5) reports the corresponding statistics from Goyal, van der Leij & Moraga-Gonzalez (2006) for the network of authors of all articles published over 1990–1999 in economics journals indexed in EconLit. Compared to the EconLit network, the IMF Working Papers network is much smaller and considerably more integrated, as evident from the much greater average degree, coverage of the giant component, and average clustering coefficient. Arguably, the most important explanation for this is that we focus on the publications of one institution, where potential collaborators may get to know each other more easily (even just because staff mostly works in the same premises) and researchers cover a smaller set of topics than in the field of economics more generally. Moreover, multi-authored publications are more common in our sample, especially so in more recent years.

Despite the IMF Working Papers network can be considered a special case of co-authorship networks, with people generally working in the same premises, the network measures in Table 1 indicate that the IMF Working Papers network exhibits "small world" properties, broadly in line with--but to an even greater degree than--what has been found for collaborations in economics journal articles (Goyal, van der Leij & Moraga-Gonzalez, 2006) and published papers and preprints in other fields of research (Newman, 2001). That is, the network consists of a large set of authors with few direct co-authors (relative to all possible collaborations); the network is nonetheless highly integrated, as the majority of authors are in some way connected to each other, mostly through short chains of co-authors; and the extent of overlap in co-authorship is very high (relative to a network where collaborations would be purely random). As Goyal, van der Leij & Moraga-Gonzalez (2006) and Hsieh et al. (2018) show, such patterns typically derive from a network architecture of interlinked "star" authors, a set of researchers that collaborate with many others, most of which are not direct co-authors themselves. These stars effectively act as connectors of different clustered parts of the network. Additional evidence for such a set-up in the IMF Working Papers network comes from the observed negative relation between individual authors' degree and their local clustering coefficient, as portrayed in Figure 2. Whereas the many researchers with only two or three co-authors have on average (at the median) around 80 percent (100 percent) of their co-authors collaborating with each other, the fewer researchers with over 20 co-authors seldom have clustering coefficients in excess of 0.2.

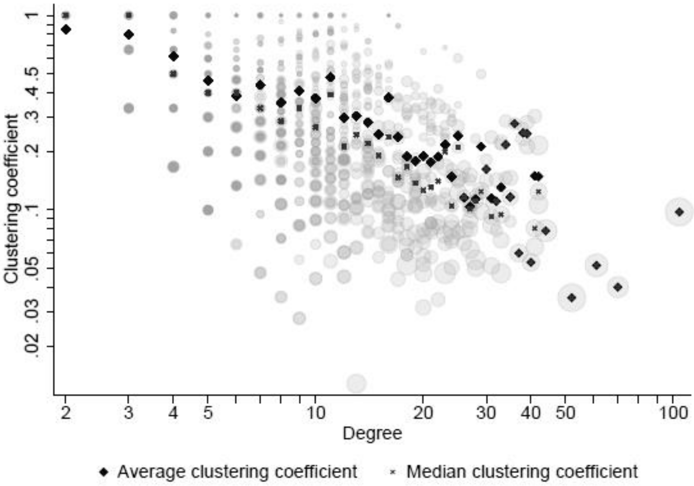


Figure 2. Author degrees vs. clustering coefficients

Notes: Dots represent individual authors, with dot size proportional to the total number of publications by author. Darker shades of grey correspond to more authors having the same degree-clustering coefficient combination.

Empirical strategy

It is reasonable to assume that researchers' decisions with whom they collaborate are mostly non-random, based on factors such as complementarity of skills, common research interests, and similarities in background. While some of these dimensions can be easily assessed from publicly available information (e.g., from researchers' publication track record), other relevant knowledge maybe only obtained indirectly through the researcher's network of acquaintances, of which the co-author network is an important subset. This would imply that, in addition to other pair and individual author characteristics, researchers' positions vis-à-vis each other in the co-authorship network may influence their decision to collaborate.

To study the determinants of first and subsequent collaborations between pairs of authors, we estimate the following logit specification on the sample of active pairs:

$$Pr(y^{i,j}_t = 1) = \Phi(\alpha + \beta N^{i,j}_t + \gamma C^{i,j}_{t-1} + \delta HR^{i,j}_t + \varepsilon^{i,j}_t)$$

where Φ is the logit function and the dependent variable $y^{i,j}_t$ is the dummy variable tracing first collaborations or, alternatively, subsequent collaborations, as defined in the Data section. The set of regressors includes network variables, $N^{i,j}_t$; pairs' characteristics, $C^{i,j}_t$ (both similar to the ones used in Fafchamps, van der Leij & Goyal, 2010); as well as our unique author-level data on IMF employment and demographics, $HR^{i,j}_t$. The variation in our dataset is mostly cross-sectional, as the large majority of author pairs were active only for a short period of time, thereby reducing the need (and feasibility) to control for pair fixed effects.

To compute the network variable regressors, we first define the network graph G_t with the authors being the nodes and the collaborations between t - 9 and t being the edges. We allow for a 10-year window as relations formed at the time of a co-authorship arguably tend to last for some time. The first network variable we include in the regressions is proximity, which is defined as the inverse of the (shortest) distance between the pair of authors in G_t. When computing this distance, however, any co-authorship between i and j is ignored, so that the minimum distance is equal to two (i.e., the pair has a common co-author) and maximum proximity is equal to 0.5 (even after the initial collaboration). This allows us to use the variable also as a potential determinant of subsequent collaborations. If there is no link between i and j, proximity takes the value of zero. The second network variable is the number of shortest paths between authors i and j in G_t, which is also computed excluding the direct link between i and j. If the two authors are not connected at all, the number of shortest paths takes the value zero.

The pairs' characteristics include research ability, the propensity to collaborate, and research (area) overlap. We proxy research ability with a measure of research productivity. More productive or talented authors are, in general, looked up to and other authors may want to connect with them. As a result, they may be more likely to start new collaborations. At the same time, relatively talented authors may be more interested in working with equally talented co-authors. Hence, we compute a productivity score for each author $i \in G_t$ as the number of "points" author i accumulated between t - 9 and t, where points for each published working paper are given by *citations*/(*number of authors* + 1). We use the average of the productivity scores for each couple of authors to get an aggregate productivity measure for the pair. We also compute the absolute value of the difference in the productivity scores of the two authors to proxy the difference in research ability. The sign of the coefficient for the latter variable will depend on whether more productive researchers look to match with each other or whether co-authorship reflects collaborations between experienced researchers and juniors. In the regression we include the one-year lag of both the average and the absolute difference in productivity to mitigate simultaneity concerns.

A researcher's propensity to collaborate is also likely to affect the probability of starting new collaborations. A researcher with many collaborators in the recent past may be intrinsically more eager to also collaborate in the future. Moreover, if an author is already well-connected, he is easier to reach out to and this may lead to new collaborations. A proxy for that propensity is the number of co-authors that each author had over the past 10 years. The measures of propensity to collaborate for the pair are the lagged average of each author's propensity to collaborate, as well as the lagged absolute value of the difference in the propensities. Research overlap is linked to shared research interests and may thus also condition the probability that two authors start a new collaboration. Researchers sharing similar interests are probably attending similar conferences, seminars, and other events, which create opportunities to connect and start collaborations. We measure research overlap as in Fafchamps, van der Leij & Goyal (2010), relying on the JEL codes we discussed in the Data section. Specifically, we define the field of research with the first letter of the JEL classification; and calculate for each author i the fraction of working papers that has been dedicated to each JEL classification letter between t - 9 and t. Working papers with multiple JEL codes are assigned proportionally to each of the different codes. We then calculate the research overlap between authors i and j using a cosine similarity function and include its lag in the regressions. For subsequent collaborations, the calculation of research overlap excludes the papers on which the authors of the pair previously worked together.

Finally, we also control for other employment and demographic characteristics. Specifically, we include a set of dummy variables indicating whether the pair of potential co-authors work in the same IMF department or did so in the past year; whether their gender is the same; and whether their region of citizenship is the same.

Results and discussion

First collaborations

We first present the results for the determinants of starting a new collaboration. The sample includes both collaborating author pairs and author pairs that never collaborated but were active. In Table 2, we report the logit regression results in terms of odds ratios. Column (1) shows that network distance is a key determinant of first collaborations. For example, for a pair of unconnected researchers (i.e., with proximity at zero) that becomes connected via a common co-author (i.e., with proximity at 0.5), the odds of starting a new collaboration are about 30 times higher; similarly, doubling the number of shortest paths connecting the researchers in the pair is associated with an increase in the probability of initiating a collaboration by a factor of nearly four. These effects are large, which is expected in the case of new collaborations.

Table 2. Determinants of first collaborations

	(1)	(2)	(3)	(4)	(5)
Proximity	6.309***	11.816***	19.365***	20.576***	16.023***
	(0.141)	(0.578)	(1.374)	(2.231)	(1.046)
Ln number of shortest paths	3.826***	12.052***	19.987***	20.078***	15.265***
	(0.149)	(0.823)	(1.709)	(2.349)	(1.108)
Lag avg. productivity		1.003	1.006***	1.007***	
		(0.002)	(0.002)	(0.002)	
Lag abs. diff. in productivity		0.996**	0.997**	0.998	
		(0.002)	(0.001)	(0.001)	
Lag avg. propensity to collaborate		0.821***	0.819***	0.795***	
		(0.006)	(0.006)	(0.007)	
Lag abs. diff. in propensity to collaborate		1.070***	1.061***	1.082***	
		(0.005)	(0.005)	(0.006)	
Lag of research overlap			1.030	0.951	
			(0.075)	(0.083)	
Same department				5.096***	
				(0.238)	
Same gender				1.007	
				(0.047)	
Same region of citizenship				1.278***	
				(0.067)	
Observations	6,040,424	3,729,804	2,374,994	1,921,742	1,921,742
Pairs	2,310,620	875,197	682,589	526,841	526,841
Pseudo R^2	0.361	0.431	0.478	0.511	0.441
AUROC	0.883	0.922	0.935	0.945	0.934

Notes: Proximity is multiplied by 10. All regressions include a constant. Heteroskedasticity and autocorrelation robust standard errors in parentheses. ***p <0.01, **p <0.05, *p <0.1.

Potential co-authors with a smaller network distance may simultaneously be closer in terms of other "non-network" dimensions of distance, such as differences in research interests, socio-cultural backgrounds, or physical locations. In what follows, we control for several proxies of these dimensions and, differently from previous papers such as Fafchamps, van der Leij & Goyal (2010), we also include controls based on employment and demographic information. However, we recognize that, even after including all controls, estimates may still reflect the influence of other unobserved factors that ultimately drive co-authorship of a working paper (e.g., interest in a very specific research question within a particular sub-field, or closeness to the same water cooler or coffee corner). Hence, strictly speaking, the results on the role of network distance for research collaborations should not be interpreted in a causal way.

In column (2), we include the pairs' characteristics. As these variables enter the specification with a lag, our sample reduces to those author pairs that remained active for at least two years in a row. By excluding "one-off researchers", this sample is likely to be more representative of the relations across authors that have a somewhat more persistent interest in research. Distance variables now have even larger odds ratios, suggesting that more research-oriented people are even more prone to rely on the network around them. The average productivity of the pair of authors is significant only when other variables are included, indicating that more productive author pairs are more likely to start collaborations. Relying on the significant coefficients in columns (3) and (4), we find that a one standard deviation increase in average productivity is associated with a 9 to 10 percent higher probability of starting a new collaboration. At the same time, when the individual productivity of the authors forming the pair diverges, the probability of starting a new collaboration declines, but the effect is not robust to the inclusion of all other determinants. A larger average propensity to collaborate--measured by the number of connections for each author of the pair--is negatively associated with the probability of starting to collaborate, suggesting that authors with many connections (generally more senior staff members with an established network, more options in terms of potential collaborations, and less time to allocate to research) may be less eager to look for new ones. However, pairs of authors having more heterogeneous propensities to collaborate are more likely to start a collaboration. One possible explanation for these results is that senior staff members are often tasked with more managerial duties compared to junior staff, who has more time for research. Therefore, these findings may to some degree reflect an efficient allocation of resources.

In column (3), we include research overlap, which turns out to be insignificant for the probability of starting a new collaboration. This finding is easier to understand when we include the employment and demographic variables in column (4). The odds of starting a new collaboration for author pairs in the same department increase by a factor of more than five, compared to author pairs of researchers from different departments. That is, authors team up because they are in the same department rather than because they dealt with similar research topics in the past. Having citizenship of the same region also favors the start of new collaborations, by a factor of about 1.3. This may reflect the importance of similar backgrounds, including native languages, and education and professional experiences, among other dimensions. In the presence of the other variables, being of the same gender is not significantly related to the likelihood of starting to work together. As shown in columns (1) to (4), the magnitude of the coefficients on proximity and the number of shortest paths varies depending on what control variables we include in the specification. However, as we add controls, the number of observations declines significantly. To understand what drives the changes in the magnitude of the coefficients, in column (5) we report the results of the regression including the same regressors as in column (1) (i.e., proximity and the number of shortest paths) but restricting the sample to the one of column (4). Our findings suggest that both the restriction of the sample and the addition of extra controls play a role, as the size of the coefficients is larger than in column (1) but smaller than in column (4).

Subsequent collaborations

Table 3 presents the estimation results when the dependent variable is the dummy variable for subsequent collaborations. The number of observations is many times smaller than for initial collaborations, because the dependent variable is defined only for author pairs that collaborated at least once. The results in column (1) suggest that proximity and the number of shortest paths--calculated excluding the direct connections between each pair of researchers--are associated with a higher probability of continuing to collaborate, but the effect is unsurprisingly much smaller than for first-time collaborations (cf. Fafchamps, van der Leij & Goyal, 2010). For example, reducing the distance between a pair of researchers from three to two almost is associated with a doubling of the chances of continuing to collaborate (compared to an increase by a factor of 11 in the case of initial collaborations); similarly, doubling the number of shortest paths connecting the researchers in the pair is associated with an increase in the probability of repeating a collaboration by a factor of about 1.5. In other words, the network distance between authors is particularly important to start a new collaboration, but once the connection is established, it becomes much less relevant for continuing to collaborate. This result is in line with the interpretation that collaboration networks may transmit important, not directly observable information about authors that helps to overcome matching frictions. Once this information is internalized through actual collaboration, the network effects strongly diminish.

Table 3. Determinants of subsequent collaborations

	(1)	(2)	(3)	(4)	(5)
Proximity	1.011	1.054**	1.040	1.094**	1.086**
	(0.022)	(0.026)	(0.025)	(0.041)	(0.035)
Ln number of shortest paths	1.507***	2.282***	2.217***	2.508***	1.920***
	(0.077)	(0.141)	(0.138)	(0.196)	(0.113)
Lag avg. productivity		1.014***	1.016***	1.012***	
		(0.002)	(0.002)	(0.002)	
Lag abs. diff. in productivity		0.992***	0.992***	0.994***	
		(0.002)	(0.002)	(0.002)	
Lag avg. propensity to collaborate		0.900***	0.919***	0.926***	
		(0.011)	(0.011)	(0.011)	
Lag abs. diff. in propensity to collaborate		1.067***	1.056***	1.054***	
		(0.007)	(0.007)	(0.008)	
Lag of research overlap			0.573***	0.898	
			(0.063)	(0.109)	
Same department				3.270***	
				(0.253)	
Same gender				1.242***	
				(0.104)	
Same region of citizenship				1.223**	
				(0.098)	
Observations	14,893	14,589	14,207	12,645	12,645
Pairs	3,325	3,289	3,244	2,865	2,865
Pseudo R^2	0.011	0.033	0.036	0.090	0.031
AUROC	0.574	0.636	0.637	0.718	0.627

Notes: Proximity is multiplied by 10. All regressions include a constant. Heteroskedasticity and autocorrelation robust standard errors in parentheses. ***p <0.01, **p <0.05, *p <0.1.

The results for the author pairs' characteristics are very similar to the ones for initial collaborations, as shown in column (2). Specifically, author pairs with a larger average productivity are more likely to continue collaborating. However, if productivity differences between the authors of the pair are large, they are less likely to continue working together. The number of established connections, which measures the propensity to collaborate, reduces the probability of continuing to collaborate. And larger differences in terms of the propensity to collaborate increase such probability. As in the case of new collaborations, these results can be explained by referring to the seniority of staff and the allocation of resources.

When we add research overlap in column (3), we find that it has a negative and significant effect on the probability of continuing to collaborate. However, once we include the employment and demographic variables in column (4), research overlap turns insignificant, while being employed in the same department and having citizenship of the same region are still important determinants of the probability of continuing to co-author, increasing the odds by more than a factor three and 1.2, respectively. Being of the same gender now leads to an increase in the likelihood of continuing to collaborate with a factor 1.2.

As for first collaborations, we re-estimate the specification in column (1) with the sample of column (4) to test whether the results are driven by the additional regressors and/or the restricted sample. The estimates reported in column (5) confirm that, again, both these factors play a role.

A commonly used statistic to assess the classification performance of logit models is the area under the receiver operating characteristic curve (AUROC). As reported in column (4) of Table 2, the AUROC for the richest specification of initial collaborations suggests that with a 94 probability the model will assign a higher predicted value to a randomly selected collaboration than to a randomly chosen non-collaboration. Such probability is somewhat lower at 72 percent in the case of subsequent collaboration, as shown in column (4) of Table 3. Overall, we conclude that our logit models have a satisfactory performance in terms of distinguishing collaborating from non-collaborating (or no longer collaborating) author pairs. We have verified that our regression results are also robust to restricting the sample to IMF Staff only; to the exclusion from the sample of all collaborations of author pairs in papers involving at least one other author that collaborated (either separately or jointly) with both authors of the pair in the past (i.e., making abstraction of "middleman/woman effects"); and to defining the network of authors and collaborations over a shorter period *t* - 4 to *t*.

Conclusion

In this paper, we study how research collaborations are structured, initiated, and maintained, using the co-authorship network of the near-universe of IMF Working Papers published during 1990-2017. We combine information on the papers with a unique dataset of demographic and employment details of the authors, which allows us to better control for possible confounders. Being a widely read research outlet where authors express their personal views on topics of their interest, and given that IMF staff does not face the typical "publish-or-perish" incentives of the academia, the IMF Working Papers series constitutes an ideal testing ground to examine the endogenous nature of co-authorship formation.

Network statistics indicate that the IMF Working Papers co-authorship network became larger and more integrated over the three decades of the sample period. Moreover, the network exhibits "small world" properties, even more so than in the broader but earlier sample of economics journal articles constructed by Goyal, van der Leij & Moraga-Gonzalez (2006): it consists of many authors with few direct co-authors, yet indirectly connected to each other through short co-authorship chains, and shows a significant degree of clustering. These properties are characteristic of a network architecture where interlinked "star" authors act as connectors of different clusters in the overall network.

The empirical investigation of the determinants of co-authorship relations reveals, above all, that two researchers that are closer to each other in the existing co-authorship network are significantly more likely to collaborate, especially so in the case of first-time collaborations. This result corresponds well with Fafchamps, van der Leij & Goyal (2010) and suggests that the network transmits important, not directly observable information about authors that helps to overcome initial matching frictions.

In addition, we find that author pairs with a higher average productivity are more likely to start and continue collaborations. At the same time, researchers with an established network of co-authors seem to be less in search of new ones; yet, authors with different co-authorship network sizes are more likely to start and maintain a collaboration. These findings likely reflect synergies between senior staff members--generally tasked with more managerial duties--and junior staff--who have more time for research. Being employed in the same department and having citizenship of the same region increase the likelihood of both starting and maintaining a collaboration. Meanwhile, greater overlap in research areas, derived from JEL codes used in earlier papers, has no significant independent effect. This suggests that similar backgrounds, including native languages and culture, more so than common research interests, play a role in the decision to team up--a novel result that we were able to uncover thanks to our rich author-level data. Finally, being of the same gender is not significantly related to the probability of initial collaboration, but once a co-authorship link is formed, authors of the same gender are more likely to continue collaborating.

Our findings provide important insights for formulating policies aimed at fostering knowledge generation and diffusion, which ultimately promote economic growth. In particular, the results of this paper suggest that if research collaboration is to be stimulated, incentives should be directed to researcher pairs that are initially more distant from each other in the co-authorship network, and that have lower average productivity, similar network sizes, and/or more heterogeneous backgrounds.

Acknowledgments and disclaimer

The views expressed do not necessarily represent those of the IMF, the National Bank of Belgium, or the Eurosystem. We thank, without implicating, Roberto Guimaraes-Filho and his team for compiling the IMF employment and demographic data.

References

Azoulay, P., Graff Zivin, J.S. & Wang, J. (2010). Superstar extinction. *Quarterly Journal of Economics*, 125(2), 549-589.

Borjas, G.J. & Doran, K.B. (2015). Which peers matter? The relative impacts of collaborators, colleagues, and competitors. *Review of Economics and Statistics*, 97(5), 1104-1117.

Fafchamps, M., van der Leij, M.J. & Goyal, S. (2010). Matching and network effects. *Journal of the European Economic Association*, 8(1), 203-231.

Goyal, S., van der Leij, M.J. & Moraga-Gonzalez, J.L. (2006). Economics: An emerging small world. *Journal of Political Economy*, 114(2), 403-412.

Hsieh, C.-S., König, M., Liu, X. & Zimmermann, C. (2018). Superstar economists: Coauthorship networks and research output. *CEPR Discussion Paper*, DP13239.

Kuld, L. & O'Hagan, J. (2018). Rise of multi-authored papers in economics: Demise of the 'lone star' and why? *Scientometrics*, 114, 1207-1225.

Liu, Y., Wu, Y., Rousseau, S. & Rousseau, R. (2020). Reflections on and a short review of the science of team science. *Scientometrics*, 125(2), 937-950.

Martin-Martin, A., Orduna-Malea, E., Thelwall, M. & Delgado Lopez-Cozar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), 1160-1177.

Newman, M.E.J. (2001). The structure of scientific collaboration networks. *PNAS*, 98(2), 404-409.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5), 71-102.

Wuchty, S., Jones, B.F. & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827), 1036-1039.

Clustering social sciences and humanities publications: Can word and document embeddings improve cluster quality?

Joshua Eykens[*], Raf Guns and Tim Engels

joshua.eykens@uantwerpen.be, raf.guns@uantwerpen.be, tim.engels@uantwerpen.be
Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, University of Antwerp,
Middelheimlaan 1, B-2020 Antwerp (Belgium)

Abstract

We study how different document representation techniques affect the outcomes of clustering based on textual information. Our dataset consists of titles and abstracts for 15,907 publications from the social sciences and humanities. We compare established document representation techniques such as TF-IDF and Latent Semantic Indexing with word and document embedding techniques (Word2Vec and Doc2Vec) in terms of the quality of the clustering outcomes. Quality is assessed by calculating silhouette scores and the average textual coherence of the clusters. The results show that word and document embeddings are promising feature engineering techniques in the context of clustering social sciences and humanities publications. The average of Word2Vec embeddings works best for identifying textually coherent clusters.

Introduction

Clustering scientific documents or publications based on relations between them has been at the core of scientometric research since the early years of the field (Small & Crane, 1979). When no or only very coarse document classification systems are in place, or when we want to develop a better understanding of the topics or specialties a document belongs to, clustering becomes an important go to method. It allows to organically construct groupings of similar documents based on specific commonalities between them from the bottom up (Waltman, Boyack, Colavizza, & Van Eck, 2020). Depending on the granularity of the final clustering results, light can be shed on the knowledge base of disciplines, subdisciplines, or research specialties (Sjögårde & Ahlgren, 2018, 2019).

Throughout the years, many clustering techniques have been studied and the appropriateness of different bibliographic variables has been analyzed. Co-citation relations between documents or journals and bibliographic coupling have traditionally been at the center of the stage. In more recent years, topic modeling and other textual relations have been explored extensively (Boyack et al., 2011; Wang & Koopman, 2017). In addition, hybrid combinations of text and citation data have proven to be promising when it comes to identifying granular clusters of documents (Boyack & Klavans, 2014; Janssens, Zhang, De Moor, & Glänzel, 2009). In this study we investigate textual features and their usefulness for document clustering. We study the clustering outcomes for a dataset which contains publications from the social sciences (SS) and humanities (H) (SSH). For a considerable share of the documents in this dataset, no citation or reference data are available. Citation coverage is a well-known problem in the context of the SSH. We compare established document representation techniques such as TF-IDF and Latent Semantic Indexing with word and document embedding techniques (Word2Vec and Doc2Vec) in terms of the quality of the clustering outcomes. Quality is assessed by calculating silhouette scores, textual coherence of the clusters, and inspecting a cluster visualization. We conclude with a discussion of the results, the limitations, and pathways for future research.

[*] Mr. Eykens is winner of the 2021 Eugene Garfield Doctoral Dissertation Scholarship Award

Data

Data are collected from VABB-SHW, the Flemish bibliographic database for the SSH (Verleysen, Ghesquière, & Engels, 2014). We have selected all journal article publications for which abstracts and titles are available for the publication years 2011–2015. Titles and abstracts of English language publications have been merged together into one text field (TAs). TAs that were shorter than 65 words have been dropped; after this operation, 15,907 publications are left in the data set. SpaCy is used to tokenize the TAs. First a TA is split into individual tokens. These tokens are lower-cased and stemmed with the Porter stemming algorithm. Finally, stop words and punctuation are removed as well as tokens shorter than 3 characters.

Methods

k-Means clustering and clustering quality

In their seminal study 'Mapping the backbone of science', Boyack and colleagues make use of the *k*-Means algorithm to cluster documents based on their relative location on different maps. In more recent work, Wang and Koopman (2017) compare the performance of *k*-Means and the Louvain algorithm in the context of their Astro dataset. They apply these two clustering algorithms on semantic representations of the articles. For their application, both *k*-Means and the Louvain algorithm appeared to be competitive with other clustering solutions (Wang and Koopman, 2017, p. 1029). For this study, *k*-Means clustering is implemented in Python with the open source machine learning library scikit-learn (Pedregosa et al., 2011). The Mini Batch variant is used (Sculley, 2010). MiniBatch *k*-Means uses randomly sampled mini-batches or 'chunks' of the data to reduce the computation time. When a sample is drawn, the *k*-Means algorithm is run and the centroids are initiated and updated.

The quality of the clustering outcomes is measured by calculating average Silhouette scores and a metric for textual coherence. The Silhouette score is a measure of internal consistency or cohesion of the clusters compared to all other clusters (the separation) (Rousseeuw, 1987). The average Silhouette score thus gives an idea of the overall performance of the clustering solution. The textual coherence is a metric based on the Jensen-Shannon divergence (JSD). JSD computes the distance between two probability distributions; in this case the word vectors for the documents and the clusters they belong to. A higher score indicates more textually coherent clusters. This metric has been described extensively in Boyack et al. (2011).

For visual inspection, we map the clustering outcomes of *k*-Means on a t-SNE plot. t-SNE or t-distributed stochastic neighbor embedding is a highly effective technique well-suited for reducing high-dimensional data to two or three dimensions (van der Maaten & Hinton, 2008). Additionally, journal discipline classifications for the publications are mapped onto the t-SNE visualization.

Term frequency and inverse document frequency (TF and TF-IDF)

For each document, the number of times a term appears is counted (Term Frequency). Term frequency-inverse document frequency (TF-IDF) is a modification of this counting scheme and corrects for very common terms, or terms which are specific to a document. IDF is the inverse function of the number of documents in which that term occurs and is multiplied with the TF. This standard approach to text vectorization has proven itself over the years and has repeatedly been shown to be a worthy competitor for more advanced NLP (Natural Language Processing) techniques (Lelu & Cadot, 2019).

Latent Semantic Indexing (LSI)

The second approach tested is Latent Semantic Indexing (LSI), or Latent Semantic Analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). LSI makes use of Singular Value Decomposition to transform the document-term matrix into a reduced matrix with a fixed number of latent 'topics' or factors as features. LSI takes into account co-occurrences of words – the semantic meaning – to reduce a document to a predefined number of topics. The technique has been shown to be successful for large scale clustering applications with scientific publications (Boyack et al., 2011). We make use of the Gensim implementation of the LSI algorithm. We have evaluated the clustering outcomes for different numbers of latent topics ranging from 50 to 1000.

Word2Vec and Doc2Vec

Word2Vec is a technique to represent words with a unique list of numbers, the vectors (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). The context of the word is used to build a semantically meaningful vector representation of a word. It is a deep learning model that tries to predict the word to be embedded by training a neural network with the words in its direct context. Documents vary in length, however, and Word2Vec is a model that transforms individual words into vectors. Each document will be represented as a vector of n terms vectors. To obtain vectors of the same dimension for each document, we first sum all n terms word vectors per document and then calculate the arithmetic mean such that each document has a single x-dimensional vector of the same length. We use the Gensim implementation of the skip-gram model with negative sampling for our analysis.
Another extension of Word2Vec is Doc2Vec (Le & Mikolov, 2014). In Doc2Vec a document is represented by a vector of fixed size. Doc2Vec trains a model to construct a unique vector for every document. These unique document vectors are concatenated with the word vectors which are shared between documents and then averaged. Each document vector is thus a combination of a paragraph or document vector and Word2Vec vectors. Doc2Vec has been implemented in the study of scientific publications and has, for example, been used in conducting detailed similarity analysis among paragraphs (Thijs, 2020).

Results

The optimization of the *k* value for *k*-Means according to the average silhouette scores led to very different results in terms of the granularity of the clustering solution. TF-IDF and LSI produced solutions of 73 and 42 clusters respectively. LSI produces clusters of higher quality in terms of silhouette scores. In contrast, the textual coherence is higher for TF-IDF. This might not be a surprise, as smaller clusters tend to be more textually coherent.

Table 1. Clustering quality for the different feature processing techniques.

	Number of clusters	Feature vector size	Average Silhouette score	Textual coherence (above random)
TF-IDF	73	9,849	0.01362	0.0422
LSI	42	50	0.0724	0.0389
Word2Vec average	40	25	0.0908	0.0434
Doc2Vec	22	15	0.0520	0.0324

The Word2Vec average strategy turned out to be the best performing feature vectorization method. With this processing technique, the *k*-Means algorithm divided the set of documents into 40 clusters. Whereas the average silhouette score is only slightly better than that for LSI, the average textual coherence is higher than was the case for other methods. This is an interesting finding, as the textual coherence is biased towards smaller clusters. As already noted by (Boyack et al., 2011), here too the textual coherence of the clusters decreases as the size increases. It would thus be no surprise if we were to find a higher coherence for the TF-IDF solution. The latter yielded 73 and generally smaller clusters. The computational cost, however, is quite large. The number of features for TF-IDF is 394 times larger than the number of features needed for Word2Vec.

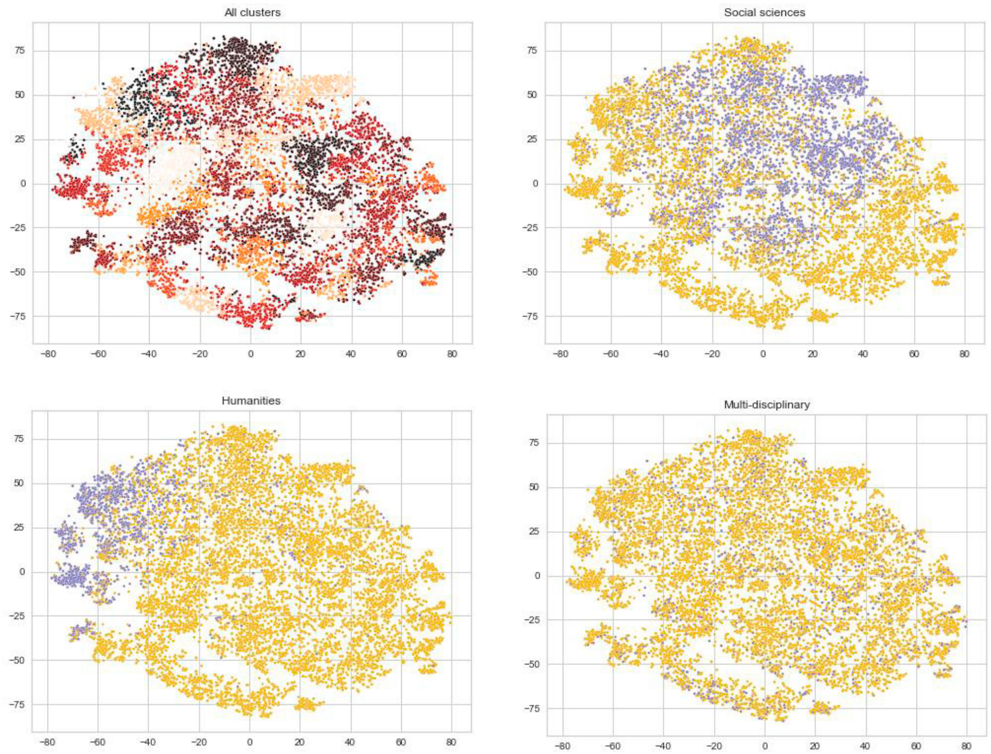


Figure 1. visualization of t-SNE embedding. (a) upper left: K-means overlay, (b) upper right: publications in SS journals highlighted, (c) publications in H journals highlighted, (d) publications in multi-disciplinary journals highlighted. Note: Axes should not be interpreted quantitatively.

In figure 1 we display t-SNE visualizations of the Word2Vec average document embeddings. The scatterplot in the upper left corner overlays the t-SNE embedding with shades indicating the clustering generated by *k*-Means. It becomes clear that, although most groups are more or less distinguishable in both cases, there exists considerable overlap between the clusters generated by *k*-Means. The remaining three plots overlay the t-SNE embedding with journal classifications of the publications. The journals have been classified according to the OECD fields of science (for details see Guns, Sīle, Eykens, Verleysen, and Engels (2018). In the upper-right corner, publications in SS journals are highlighted. We can see that the upper-right section of the scatter plot is densely populated with publications in SS journals, gradually fading to the upper left.

On the scatterplot in the lower-left corner publications in humanities journals are highlighted. Here we see that they are positioned opposite to the publications in SS journals highlighted in the upper-right scatterplot. The upper left region of the scatterplot highlighting the humanities seems to be most densely populated, fading to the right hand side, the corner in which SS publications are. This fading over is an interesting pattern, indicating that there might be important cognitive cross-over areas between both the SS and H. As we can see in the upper left corner, one of the clusters which is found by *k*-Means (in white) nicely overlaps this fading area. The fourth and last scatter plot highlights publications in multi-disciplinary journals.

These are journals which have been assigned to multiple OECD FOS discipline codes. As one would expect, these publications are dispersed over the scatterplot, popping up in multiple clusters of documents.

Limitations

The clustering algorithm used in this study demands for a pre-specified number of clusters. Additionally, only one similarity measure has been tested (Euclidean distance). It would be interesting to compare the *k*-Means clustering algorithm as well as the similarity measure to other current approaches used in a bibliometric context. Boyack et al. (2011), for example, make use of the DBSCAN algorithm and cosine similarity. Other studies make use of the Leiden algorithm or the Louvain algorithm for community detection (Wang & Koopman, 2017). Similarity approaches like the cosine similarity and BM25 similarity have been shown to perform well in a text clustering context (Waltman et al., 2020).

Conclusion and future research

In this paper we have shown that state-of-the-art vectorization techniques (Word2Vec and Doc2Vec) work well in the context of clustering SSH publications based on textual information. Although the outcomes of the different vectorization techniques in terms of silhouette scores and textual coherence were nearly identical, Word2Vec average turned out to be the best strategy to identify 40 well-divided and coherent clusters. The contextual sensitivity of Word2Vec might be a possible explanation.

A next step consists of studying different clustering algorithms and similarity calculations. Additionally, studying these clusters of documents in detail will yield interesting insights into specialized communities. The clusters can be thought of as the knowledge base of research specialties or disciplines. In another phase we will add an additional, more granular layer to the clustering presented above. A detailed content analysis and second round of clustering for the different document groups will yield insight into what these clusters actually represent. Are they representations of one or many subject specialties?

We hypothesize that while some document clusters might be representative of local or regionally oriented specialties, others will be part of the knowledge base of more global communities. For the latter, it will be difficult to reach any conclusions without taking the broader context into account. For the former, however, by studying their bibliographic, cognitive and social characteristics we will be able to broaden our understanding of research specialties in the SSH as well as bibliographic units.

Bibliography

Boyack, K. W., & Klavans, R. (2014). Including cited non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics, 8*(3), 569-580. doi:10.1016/j.joi.2014.04.001

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., . . . Börner, K. (2011). Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. *PLoS ONE, 6*(3), e18029. doi:10.1371/journal.pone.0018029

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6). doi:https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Guns, R., Sīle, L., Eykens, J., Verleysen, F. T., & Engels, T. C. E. (2018). A comparison of cognitive and organizational classification of publications in the social sciences and humanities. *Scientometrics, 116*(2), 1093-1111. doi:10.1007/s11192-018-2775-x

Janssens, F., Zhang, L., De Moor, B., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management, 45*(6), 683-702. doi:10.1016/j.ipm.2009.06.003

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning* (Vol. 32). Beijing, China: JMLR: W&CP.

Lelu, A., & Cadot, M. (2019). Evaluation of text clustering methods and their dataspace embeddings: an exploration. In *IFCS 2019 - 16th International on the Federation of Classification Societies* (pp. 1-24). Thessaloniki, Greece.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems* (Vol. 2, pp. 3111-3119).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*(85), 2825-2830.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

Sculley, D. (2010). Web-scale k-means clustering. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web* (pp. 1177-1178).

Sjögårde, P., & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics, 12*(1), 133-152. doi:10.1016/j.joi.2017.12.006

Sjögårde, P., & Ahlgren, P. (2019). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties. *Quantitative Studies of Science, 1*. doi:10.1162/qss_a_00004

Small, H., & Crane, D. (1979). Specialties and disciplines in science and social science: An examination of their structure using citation indexes. *Scientometrics, 1*, 445-461.

Thijs, B. (2020). Using neural-network based paragraph embeddings for the calculation of within and between document similarities. *Scientometrics, 125*, 835-849. doi:https://doi.org/10.1007/s11192-020-03583-6

van der Maaten, L., & Hinton, G. (2008). Visualising Data using t-SNE. *Journal of Machine Learning Research, 9*(86), 2579-2605.

Verleysen, F. T., Ghesquière, P., & Engels, T. C. E. (2014). The objectives, design and selection process of the Flemish Academic Bibliographic Database for the Social Sciences and Humanities (VABB-SHW). In W. Blockmans, L. Engwall, & D. Weaire (Eds.), *Bibliometrics. Use and abuse in the review of research performance* (pp. 117-127). London: Portland Press.

Waltman, L., Boyack, K. W., Colavizza, G., & Van Eck, N. J. (2020). A principled methodology for comparing relatedness measures for clustering publications. *Quantitative Science Studies, 1*(2). doi:https://doi.org/10.1162/qss_a_00035

Wang, S., & Koopman, R. (2017). Clustering articles based on semantic similarity. *Scientometrics, 111*(2), 1017-1031. doi:10.1007/s11192-017-2298-x

Status and Challenges of Open Government Data in Nigeria: An Informetric Analysis of Websites of Government Ministries and Organizations

Ifeanyi J. Ezema

ifeanyi.ezema@esut.edu.ng
University Librarian, Enugu State University of Science and Technology Enugu (Nigeria)
Department of Information Science, University of South Africa (South Africa)

Abstract

Globally, democratic principles revolve around transparency, accountability and greater involvement of the citizens in government programmes and activities for sustainable development. This can only be achieved if the people have access to government information and this is why agitations for freedom of information continue to heighten. Open government data (OGD) play major roles in democratization of information and a propeller of transparency and open governance. This paper adopted descriptive informatics and literature review approach to examine the websites of major government ministries and agencies in Nigeria including Nigeria Data Portal to determine the availability and status of open government data. Findings reveal that while some websites have presence of OGD, over 60% are yet to provide such data in the website. Also the data portal has several government data which are rarely updated. The paper also identified several challenges of OGD in Nigeria and proposed recommendations to mitigate the challenges.

Introduction

The major ingredients of democracy are citizens' active participation and transparency of leaders in the activities of government through open governance. Agitations for democratic culture globally is often anchored on human rights which entails freedom of expression, rights to government information and ability to use these rights for sustainable socio-economic and political development. Since Nigeria returned to democratic government in 1999, citizens have been deeply engaging government for good leadership anchored on open and good governance. An interesting way of achieving open government is making government data available to the people (that is freedom of the citizen to access government information). Fortunately, Nigeria enacted Freedom of Information Act in 2011, (Federal Republic of Nigeria 2011) and by implication, Nigeria government is expected to run an open government driven by open government data (OGD) which will facilitate wider access to public sector information (PSI).

Open data is defined by The International Open Data Charter (ODC) as "publicly available data that can be universally and readily accessed, used and redistributed free of charge. It is structured for usability and computability" (Van-Belle et.al. 2018). But Attard, Orlandi, Scerri, & Auer, (2015: 402) defined Open Government Data as "government-related data that is made open to the public." This appears similar to that of Economic Commission for Africa (2017) which described open government data as "any government-produced or government-commissioned data that are freely available and publicly accessible." Perhaps, a more elaborate definition is that provided by Janssen & Helbig (2015) and Jetzet (2016) who observed that such data is multi-dimensional providing valuable information for socio-economic and technological development of the country. Global interest in OGD began as a result of Open Government movement within the first decade of 21st century (Bauer & Kaltenbock 2012). Open government data and open governance are interrelated because it is through the availability of open government data that a country can achieve open government.

Interest in open government data is heightened by its critical contributions to national development and impact in transparent governance. The fruits of such open governance are opportunities for speedy economic growth (Jetzek, Avital, & Bjorn-Andersen, 2012), public accountability (Viscusi, Spahiu, Maurino, & Batini, 2014; Worthy, 2015); contributions to social inclusion (World Wide Web Foundation 2015), job creation and growth of knowledge economy (Carrara, Chan, Fisher & van Steebergen 2015). Vickery (2011) provided economic implications of open government data estimating the aggregate direct and indirect economic impacts from use of public sector information in the European Union economy to the tune of billions of euros yearly.

Recent developments reveal a growing interest in the public use of OGD and also from the academia in the reuse of government data for academic and research purposes and this has been underscored by the study of Kučera (2018) who posited that globally OGD forms critical component of research material since government has the infrastructure and personnel for aggressive data collection not withstanding several challenges. Such challenges to a large extent impede the implementation of freedom of information in many African countries as reported by Asogwa & Ezema (2017). The publication of OGD is another challenge and has been well documented in the works of Chlapek at al., (2014); Kucera (2015) and Kucera et al., 2015).

A recent study by Ezema (2019) examined open government data in selected government websites in Nigeria and revealed an availability of open government data in six of the eight government agencies selected for the study. However, a major weakness of the study is that the number of government agencies covered is so small and requires further investigation. With the growth of electronic governance, government data continue to emerge daily in several government websites, but little is known about the status of these data in Nigeria; their management for research and scholarly communication; and likely management challenges – as major studies on OGD in Africa did not cover Nigeria (see Economic Commission for Africa (2017) & Van-Belle et. al. 2018). The purpose of this paper therefore, is to examine the status open government data in Nigeria using informetric analysis of government websites, explore their management challenges and provide recommendations for improvement. Specifically, the paper intends to:

a. Ascertain the availability of open government data in major government agencies in Nigeria
b. Determine the adoption and availability of open government data in Nigeria government ministries
c. Examine the presence of open government data in Nigeria data portal
d. Determine the currency of the available open government data in Nigeria data portal.
e. Explore the challenges of managing open government data in Nigeria

Materials and Methods

This study adopted descriptive informetrics to examine some government websites in Nigeria so as to determine their status in terms of presence of open government data, their types, currency and accessibility. A literature review approach was also used to articulate strategies of managing them and challenges therein. Data were extracted from the websites of 26 Federal Government Ministries and 16 selected major government agencies and parastatals (from September 2 to 10 2020) to determine the availability and status of open government data in the websites. Within the same period additional, data were extracted from Nigeria Data Portal http://nigeria.opendataforafrica.org/ which is maintained by the National Bureau of Statistics responsible for collection of data on each of the 36 states in Nigeria and Abuja. The portal covers many key areas of national development such as health, agriculture, education, trade and investment, women and child issues, demography among others. Data from the websites were analyzed by identification of presence of open government data with a remark on the expected data to be seen in the websites, data from Nigeria Data Portals were analyzed using frequency and percentages and presented in tables.

Results

Table 1: Selected Government agencies in Nigeria and availability of Open Government Data

S/N	Name of Government Agency	Websites	OGD presence	Remarks
1	Budget and National Planning	http://www.nationalplanning.gov.ng	N/A	Budget proposal and breakdown. Real data not available
2	Central Bank of Nigeria	https://www.cbn.gov.ng/	A	Data on GDP, inflation and other financial issues
3	Economic and Financial Crime Commission	http://www.efccnigeria.org/efcc/	N/A	Data on convictions. No data on recoveries made No data on state corruption index
4	Independent National Electoral Commission	https://www.inecnigeria.org/	A	Data on only recent election results Absent of data on past election result
5	National Bureau of Statistics	https://nigerianstat.gov.ng/	A	Robust data all regions of Nigeria but many of them not updated.
6	National Population Commission	http://population.gov.ng	N/A	No active website at the time of study
7	Nigeria National Petroleum Corporation	https://www.nnpcgroup.com	A	Data on oil and gas production and trade
8	Nigeria Stock Exchange	http://www.nse.com.ng/	A	Robust data on financial market
9	Independent Corrupt Practices and Other Related Offences Commission	https://icpc.gov.ng	N/A	Annual report up to 2015 No data on cases investigated No data on recoveries made No data on states corruption index
10	Nigerian Centre for Disease Control	https://ncdc.gov.ng	A	Data on weekly epidemiological report from 2016. A-Z list of infectious diseases with their symptoms Disease situation report Update on Covid-19 in Nigeria
11	National Agency for Food and Drug Administration and Control	https://www.nafdac.gov.ng/	N/A	There is a page on drug data but no content No periodic data on registered drugs No periodic data on drug abuse and rehabilitation, drug reactions No data on pharmaceutical industries
12	National Agency for the Prohibition of Trafficking in Persons	https://www.naptip.gov.ng/	A	Few data on offenders. No periodic data on trafficking trends, conviction, rehabilitation of victims
13	Nigerian Communications Commission (NCC)	https://www.ncc.gov.ng	A	Presence of period statistics on connected and active telephone lines, VoIP, fixed wired/wireless from 2017 to 2020. Tele density statistics, subscribers' operation data etc.

Table 1 (cont.): Selected Government agencies in Nigeria and availability of Open Government Data

S/N	Name of Government Agency	Websites	OGD presence	Remarks
14	National Drug Law Enforcement Agency (NDLEA)	https://ndlea.gov.ng	N/A	No periodic data on drug use, drug manufacturing and distribution pattern, drug crimes and convictions.
15	National Universities Commission	https://www.nuc.edu.ng	N/A	No periodic statistics on professor/students ratio, students enrolments, graduation, academic and non-academic staff, accreditation performance, university ranking, etc.
16	Nigeria Immigration Service	https://immigration.gov.ng/	N/A	No periodic data on immigrants' registration, processed passports, visas issued, ECOWAS travel certificate, resident permits.

The table above shows that of the 16 government agencies, only 8(50%) have robust data. Surprisingly, the two agencies of government (Economic and Financial Crime Commission (EFCC) and Independent Corrupt Practices and Other Related Offences Commission) which are responsible for investigation and prosecution of financial crimes have no data in their websites to show the periodic data on arrests and number of convictions. One also expects to see data on recovered assets among other related issues. It is also a source of worry that National Population Commission has no active website considering its critical roles in overall national development.

Table 2: Availability of open government data in Nigeria Federal Ministries

S/N	Ministries	Websites	Expected Data	OGD presence
1	Agriculture	http://fmard.gov.ng/	Publications from the ministry available	N/A
2	Aviation	http://aviation.gov.ng/	Inactive website	N/A
3	Defence	http://www.defence.gov.ng/	Articles, speech and few publications are available	N/A
4	Education	http://www.education.gov.ng/	Data on number of schools, enrolment, graduate outputs etc, but data are not updated	Available
5	Environment	http://environment.gov.ng/	Data on pollution, oil spillage, open defecation etc.	N/A
6	Federal Capital Territory	http://fcda.gov.ng/	Housing development, public infrastructure, urban migration etc.	N/A
7	Finance and Economic Development	http://www.finance.gov.ng	Budget allocation, economic growth, financial institutions, trade and investment etc	Available
8	Foreign Affairs	http://foreignaffairs.gov.ng/	Nigeria's international collaborators, trade partners, foreign policies, population in diaspora etc.	N/A (Inactive website)
9	Health and Social Services	http://health.gov.ng/	Disease burden (communicable and non-communicable), health facilities, immunization trends, number of health personnel, mortality rate, covid-19 etc.	A data only on health facilities

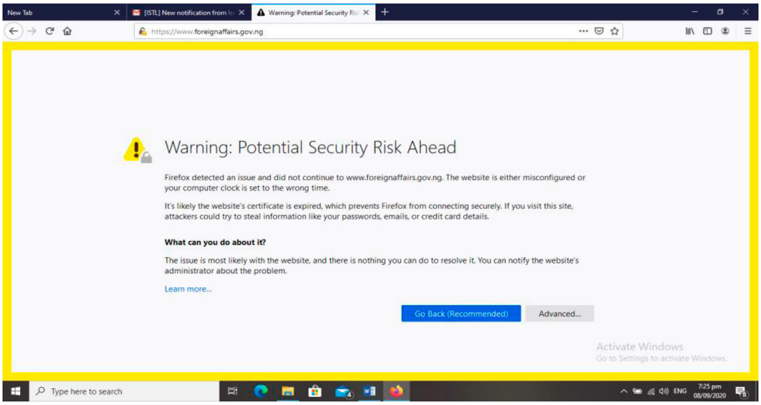
Table 2 (cont.): Availability of open government data in Nigeria Federal Ministries

S/N	Ministries	Websites	Expected Data	OGD presence
10	Information & Communication	http://fmic.gov.ng/	Data on journalists in Nigeria, statistics print and electronic media in Nigeria, household internet and telephone connectivity, etc	N/A
11	Internal Affairs	http://www.interior.gov.ng	Population statistics, death and birth statistics, immigration data, prison data, crime statistics etc	N/A
12	Justice	http://www.justice.gov.ng/	Statistics of legal personnel, data on asset recovery, jurisprudent data, statistics of bills, etc	N/A
13	Labour and Productivity	http://labour.gov.ng	Unemployment statistics, data on job creation, labour migration, child labour statistics, etc.	N/A
14	Power and Steel	https://www.power.gov.ng/	Data on power generation and distribution, households/communities connected to electricity, annual steel production and marketing etc,	Available data on power generation and forecast
15	Solid Minerals Development	https://portal.minesandsteel.gov.ng/	Data on different solid minerals, statistics of mine sites, data on geological professionals in the country, annual production of solid minerals, etc.	N/A
16	Industries	http://nid.fmiti.gov.ng/	Data about industrial growth, Nigeria's key investment both local and international, data on small and medium enterprise etc.	N/A
17	Culture Tourism and National Orientation	*www.fmtc.gov.ng*	Statistics of media houses, tourist sites, statistics of cultural festivals in Nigeria, data on hotels and related organizations, etc.	N/A
18	Niger Delta	http://nigerdelta.gov.ng/	Data on oil spillage, statistics on gas flaring, data on infrastructural development of Niger Delta region, etc.	N/A
19	Petroleum Resources	http://petroleumresources.gov.ng/	Data on petroleum production and use, statistics of oil spillage, data on international oil market, etc.	N/A
20	Works and Housing	http://www.pwh.gov.ng	Data on road networks, statistics of housing development, data on electric supply and house hold utility, statistics of building collapse, etc	N/A
21	Science and Technology	https://scienceandtech.gov.ng/	Patents statistics, scientific research data, statistics of technological innovations, etc.	N/A
22	Trade and Investment	http://nid.fmiti.gov.ng	Statistics of small and medium enterprise, export and import statistics, statistics of industries in the country, etc.	N/A
23	Transportation	http://www.transportation.gov.ng	Data on road use, statistics of air travels, ships and ferries, statistics of registered automobiles, etc.	N/A

Table 2 (cont.): Availability of open government data in Nigeria Federal Ministries

S/N	Ministries	Websites	Expected Data	OGD presence
24	Women Affairs and Social Development	https://www.womenaffairs.gov.ng/	Data on women empowerment, statistics of female enrolment in schools, data on gender equality, statistics of female genital mutilation, data on widowhood practices.	N/A
25	Water Resources & Rural Development	https://waterresources.gov.ng/	Data on household water supply, irrigation development, country's water mapping, fresh water ecosystem, aquatic life etc.	N/A
26	Youth and Sports	https://youthandsport.gov.ng/	Employment statistics, data on job creation, data on youth empowerment, etc	N/A (Inactive website)

Table 2 shows that data are available in only 4 (15.4%) of the twenty-six government ministries in Nigeria. These ministries are Education, Finance, Power and Health while the remaining 22 (74.6%) have no open government data in their website. It is even more surprising that two of the ministries (Youths and Sports Development and Foreign Affairs) have no functional websites. Also the Ministry of Youth and Sports Development with the published link https://youthandsport.gov.ng/ cannot connect to any server as at 7.53 pm September 8, 2020. For the Ministry of Foreign Affairs which portrays the country to the global community, a screen shot of the current state of the website is presented as fig. 2.


New Tab
[ISTL] New notification from I
Warning: Potential Security Ri
https://www.foreignaffairs.gov.ng
Go Back (Recommended)
Advanced...
Activate Windows
Go to Settings to activate Windows.
Type here to search
ENG
7:25 pm
08/09/2020

Warning: Potential Security Risk Ahead

Firefox detected an issue and did not continue to www.foreignaffairs.gov.ng. The website is either misconfigured or your computer clock is set to the wrong time.

It's likely the website's certificate is expired, which prevents Firefox from connecting securely. If you visit this site, attackers could try to steal information like your passwords, emails, or credit card details.

What can you do about it?

The issue is most likely with the website, and there is nothing you can do to resolve it. You can notify the website's administrator about the problem.

Learn more...

Fig 1: A screen shot of Ministry of Foreign Affairs website when visited on September 8, 2020 by 7.27pm.

To determine the status of other open government data, Nigeria Data portal (http://nigeria.opendataforafrica.org/) was visited and indications from table 3 reveal that the portal has only 181 data sets in 18 sectors with 17 unclassified data sets out of which economic sector (banking and financial management) has the highest proportion (39.2%) followed by issues related with governance (13.3%), oil and gas has 12 (6.6%) data sets. Labour and productivity has 10 (5.5%) data sets while agriculture, health and education has 7 (3.9%) each; while mining and solid minerals, water resources and disaster control has 0.6% respectively

Table 3: Open Government Data found in Nigeria Data Portal

S/N	SECTORS	No of Data sets	Percentage
1	Banking and Financial Management	71	39.2
2	Governance	24	13.3
3	Others (unclassified)	17	9.4
4	Oil and Gas	12	6.6
5	Labour and Employment	10	5.5
6	Agriculture	7	3.9
7	Education	7	3.9
8	Health	7	3.9
9	Transportation	5	2.8
10	Population	4	2.2
11	Power and Energy	4	2.2
12	Infrastructure	3	1.7
13	Crime control	2	1.1
14	Youth Development	2	1.1
15	Telecommunication	2	1.1
16	Mining and Solid Mineral	1	0.6
17	Water resources	1	0.6
18	Disaster control (Fire, flooding etc)	1	0.6
19	Land use	1	0.6
	Total	181	100.0

A major concern about the data sets found in the portal is that greater a percentage of the data sets appears to be outdated. As can be seen in table 3, analysis of a sample of major data sets from critical sectors of the government reveal that many of them have not been updated for a very long time and therefore, may be unreliable. For instance, 17(28.3%) of the data sets were last updated in 2014, while 14(23.3%) were updated last in 2016 and 10(16.7%) were updated last in 2017. Those that may be regarded as current were updated in the year 2019 and 2020 representing 11.7% and 3.3% respectively.

Table 4: Currency of the Data sets by years

Year	No of data sets	%
Year 2012	3	5.0
Year 2014	17	28.3
Year 2015	4	6.7
Year 2016	14	23.3
Year 2017	10	16.7
Year 2018	3	5.0
Year 2019	7	11.7
Year 2020	2	3.3
Total	60	100.0

Discussions

The availability of open government data in websites of government agencies in Nigeria appears to be grossly inadequate and does not present the country as prepared for running an open and inclusive government. Findings indicate that only half of them provide open government data capable of creating access to information in the website even when there is freedom of information law in the country. This appears antithetical to the potential benefits of open government as has been articulated by Hassan & Twinomurinzi (2018). It is also sad to note that some government agencies with presence of OGD, very critical data were not provided. For instance, the Independent National Electoral Commission (INEC) provides data only on recent election results, without the archives of the past ones. Similarly, data on voters' registration is scanty as it provides only the number of registered voters in Nigeria without a breakdown by states, local governments and possibly at ward levels. Also the website of National Agency for Prohibition of Trafficking in Persons contains few data on offenders with absence of periodic data on the trends of trafficking, and yearly convictions across the country; and complete absence of data on rehabilitation of victims of the traffickers. Available data provided by Nigeria Communication Commission is quite commendable as it contains adequate data on a number of variables on telecommunication in Nigeria. However; a more worrisome situation is a complete absence of data from National Universities Commission, the regulatory agency for all the universities in Nigeria.

In relation with an earlier study conducted by Ezema (2019), it can be deduced that availability of OGD dropped from 75% to 50% as more government agencies were included in the current study, but agencies without active website remain the same. This means that National Population Commission an agency with a very important task of conducting and updating the national population is still without active website at the time of data collection and this has very serious implications as most of policy issues and government decisions hinges on it.

An examination of government ministries shows even a worse situation than the government agencies discussed above. Only four of the ministries provided OGD in their websites while some of the ministries do not have functional website – an indication of absence of online digital contents for the public. A more disturbing situation is the case of Ministry of Foreign Affairs which is the mirror of Nigeria to the global community but without functional website at the time of data collection and this is also the case of Ministry of Youths and Sports Development. Policy-related issues such as merging and de-merging of some ministries by subsequent administrations in Nigeria could have resulted to this. Bu surprisingly, an attempt to visit the website of Ministry of Lands and Urban Development which has been merged with other ministries by the present Buhari Administration using the published link http://www.landsandhousing.gov.ng came up with Ministry of Justice and the two ministries have never been merged which shows some other issues rather than merging may be responsible for this development. The poor presence of OGD in the websites could also be attributed to lack of awareness of its potential contributions in driving socio-political economy of the country which has been pointed out by Helbig & Birkhead, (2015) and Kvamsdal (2017).

Interestingly, the creation of online data portal for aggregation of OGD in Nigeria is a good development as this will facilitate greater access among the public (see Nigeria Data Portal http://nigeria.opendataforafrica.org/). The portal however, requires periodic update of data because the latest population data in the portal was 2006. Other government agencies also maintain website that provide open government data for the interest of the citizens. A good example is National Bureau of Statistics (https://nigerianstat.gov.ng/) which provides open government data on key national development indicators.

Challenges of OGD in Nigeria

A major challenge of management and utilization of OGD all over the world is lack of awareness of their existence. This has been reported in the study conducted by Martin, Helbig & Birkhead, (2015) using survey and focus group discussion. Another problem generally in Africa is lack of true commitment for transparency and open government which is evident in paucity of policy framework in many African countries. Other challenges of OGD as identified by Nayek (2018) include lack of metadata standards for better discoverability and interoperability as published data often do not relate to data expected by users. These are global challenges but there are localized ones in Africa with special reference to Nigeria as summarized below:

Poor ICT infrastructure to Support Digital Open Government Data: The reality is that implementation of OGD is dependent on robust and efficient ICT infrastructure and regular power supply. Unfortunately, Nigeria still suffers from technological challenges arising from poor and irregular Internet connectivity among majority of the citizens and even government organizations (Ezema 2013). In addition, quantity of public power supply in the country is low because installed capacity for power generation is 12,522 MW but only 3,384 MW is generated to serve over 200 million people (USAID 2019) and this is grossly in adequate.

Lack of skilled personnel for OGD: A related challenge is the paucity of skilled manpower for organization and management of OGD in Nigeria. Though, studies show a relative improvement in skilled manpower to maintain technological facilities (Ezema, Ugwuanyi & Ugwu 2014), there is still need to train personnel in critical ICT areas.

Attitude towards transparency and open government*:* Though many African countries including Nigeria have enacted freedom of information laws, government programmes and activities are often shrouded in secrecy (Asogwa and Ezema 2017) and this is still prevalent in many government agencies even in Nigeria. Consequently, many government agencies and establishments find it difficult to openly publish government data and this is often the case with data that are likely to expose corrupt practices among government officials.

Policy framework for open government data: As has been observed, Nigeria's policy framework on OGD is still in a draft stage and therefore, there is no legal guideline for its implementation. It is surprising that over six years after the draft, nothing has been done with its publications for purpose of implementation.

Strategies for Managing Open Government Data in Nigeria

In Africa, Nigeria remains one of the countries that is yet to fully implement open government data. The number of countries in the region releasing their national data catalogues keeps increasing, with eight out of the 10 countries maintaining a reference catalogue of some kind (Van-Belle et al. 2018). For Nigeria to join other countries in Africa that has fully implemented OGD, the following strategies are hereby proposed:

- As major information providers to researchers and scholars, academic and research libraries are positioned to create awareness about OGD through seminars and workshops within the universities and research institutes. In addition, online selective dissemination of information where librarians provide links of government data to researchers in accordance with their research profiles will likely stimulate the interest of researchers on OGD.

- Government should as a matter of urgency publish OGD draft policy to facilitate its implementation in Nigeria. Such policy document is very important as issues such as legislations, action plan, sanctions for violation of the law and the kind of government data that must be made freely available will be well articulated.
- Government agencies should seek the services of librarians and other information professionals skilled in creation and assigning of metadata to assist in providing metadata for every open government data uploaded for public consumption. For the standardization of OGD metadata, the Nigeria Library Association should constitute a team of skilled experts to develop a standard for metadata description of open government data in Nigeria.
- To ensure greater access to open government data in Nigeria, academic and research libraries must engage their skilled staff in data mining and archiving of OGD for enhanced information services delivery to their readers. Downloaded open government data should be archived in the library websites and also provided in other electronic databases of the library.
- There is a need to link all to Nigeria data portal to create a webometric control of the portals. This should be part of the policy framework for OGD in Nigeria. There ought to be a government agency with legal mandate to collate organize and manage all open government data to ensure global visibility and accessibility to all citizens of the country.

Conclusion

It appears that there is genuine interest in implementation of open government data in Nigeria given the desire of the citizens to participate actively in the programmes and activities of government. The attempts by a few government agencies and organizations in making government data freely available on their websites are indications that the agitations for transparency and accountability are perceived outcomes of democratic culture in Nigeria. However, available evidences show that a greater proportion of government agencies are yet to provide OGD in their websites. Part of the reason perhaps is the inability of government to publish the draft policy guidelines for open government data which is an indication of lack of political will for implementation of OGD in Nigeria. The global rating of any government is dependent on level of transparency and accountability which are often assessed by free availability of open government data to the general public. This paper has been able to articulate strategies to mitigate these identified challenges against the full implementation and management of OGD in Nigeria.

References

Asogwa, B.E & Ezema, I. J (2017). Freedom of Access to Government Information in Africa: Trends, Status and Challenges. *Records Management Journal,* 27(3), 318-338, https://doi.org/10.1108/RMJ-08-2015-0029

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, *32*(4), 399–418. https://doi.org/10.1016/j.giq.2015.07.006

Bauer, F., & Kaltenböck, M. (2012). *Linked Open Data: The Essentials*. Vienna: edition mono/monochrom.

Carrara, W., Chan, W.S., Fischer, S., & van Steenbergen, E. (2015a). *Creating Value through Open Data: Study on the Impact of Re-use of Public Data Resources*. Retrieved from https://www.europeandataportal.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf. on June 18 2019.

Chlapek, D., Klímek, J., Kučera, J. & Nečaský, M., (2014). Otevřená a propojitelná data – metodiky, postupy, nástroje a praxe [Linked Open Data – methodologies, approaches, tools and the current practices]. In: Šaloun, P. & Chlapek, D. (eds.). *DATAKON 2014*. Ostrava: Vysoká škola báňská-Technická univerzita Ostrava, 2014, pp. 17–37

Economic Commission for Africa (2017). *Unlocking the potential of open government in Africa: Policy, legal and technical requirements for open government implementation in Africa.* Addis Ababa: Economic Commission for Africa.

Ezema, I.J. (2013) Local Contents and the Development of Open Access Institutional Repositories in Nigeria University Libraries: Challenges, Strategies and Scholarly Implications. *Library Hi Tech*, 31(2), 323 – 340. Available at www.emeraldinsight.com/0737-8831.htm. Accessed July 4, 2019.

Ezema, I.J (2019). Management of Open Government Data in Nigeria Academic Libraries: Status, Challenges and Strategies. Paper presented in IFLA WLIC Satellite meeting, Deutsche Nationalbibliothek (DNB), Frankfurt, Germany, 22-23 August 2019. Retrieved on January 4, 2020 from http://library.ifla.org/id/eprint/2748

Ezema, I.J, Ugwuanyi, C.F & Ugwu, Cyprian I (2014). Skills requirements of academic librarians for the digital library environment in Nigeria: a case of University of Nigeria Nsukka. *International Journal of Library and Information Science.* Available at www.iaeme.com/ijlis.asp

Federal Government of Nigeria (2014). *Nigeria's Draft Open Data Guidelines.* Retrieved on July 8, 2019 from https://docs.google.com/document/d/1Ssbsj-eTEVUcFITnus-hPMFjTgkHQT_Sypbb8UCGQLk/edit

Janssen, M., Charalabidis, Y. & Zuiderwijk, A., 2012: Benefits, Adoption Barriers and Myths of Open Data and Open Government, *Information Systems Management* 29(4), 258-268

Jetzek, T. (2016). Managing complexity across multiple dimensions of liquid open data: The case of the Danish Basic Data Program. *Government Information Quarterly*, *33*(1), 89–104. https://doi.org/10.1016/j.giq.2015.11.003

Jetzek, T., Avital, M, & Bjorn-Andersen, N. (2012). The Value of Open Government Data: A Strategic Analysis Framework Research-in-Progress. Retrieved from http://openarchive.cbs.dk/bitstream/handle/10398/8621/Jetzek.pdf?sequence=1

Hassan, M.I.A & Twinomurinzi, H (2018). A systematic literature review of open government data research: challenges, opportunities and gaps. Paper presented at the IEEE conference, October 2018. DOI: 10.1109/OI.2018.8535794

Kucera, J (2015). Open Government Data Publication Methodology. *Journal of System Integration*, 2015,2. DOI: 10.20470/jsi.v6i2.231

Kucera, J. (2018). Analysis of barkers to publishing and re-use of open government data. Retrieved on July 12, 2019 from https://idimt.org/wp-content/uploads/2017/03/2017-Topic-K-Kucera-keynote-public-draft.pdf

Kucera, J., Chlapek, D., Klímek, J. & Nečaský, M., 2015: Methodologies and best practices for Open Data publication. In: Nečaský, M., Pokorný, J., Moravec, P. (eds.). *Proceedings of the Dateso 2015 Workshop*. Praha: MATFYZPRESS, 2015, pp. 52-64

Kvamsdal, P (2017). Open Government Data - A Literature Review and a Research Agenda. Retrieved from https://www.researchgate.net/publication/326160800 on June 17, 2019.

Martin, E.G, Helbig, N, Birkhead, G.S (2015). Opening health data: what do researchers want?

Early experiences with New York's open health data platform. *Journal of Public Health Management Practice,* 21(5), E1–E7. doi: 10.1097/PHH.0000000000000127

Nayek, J.K (2018). Evaluation of open data government sites: a comparative study. *Library Philosophy and Practice* (e-journal). 1781. Retrieved from https://digitalcommons.unl.edu/libphilprac/1781 on June 12, 2019.

USAID (2019). *Nigeria: Power Africa fact sheet*. Retrieved from https://www.usaid.gov on July 13, 2019.

Van-Belle, Lämmerhirt, D, Iglesias, C, Mungai, P, Nuhu, H, Hlabano, M, Nesh-Nash, T & Chaudhary, S (2018). *African data revolution report 2018: the status and emerging impact of open data in Africa.* Retrieved on July 8, 2019 from https://webfoundation.org/docs/2019/03/Africa-data-revolution-report.pdf

Vickery, Graham, 2011: Review of recent studies on PSI re-use and related market developments. In: *Europe's Information Society Thematic Portal* [online]. Retrieved on January 2, 2021 from: http://ec.europa.eu/information_society/policy/psi/docs/pdfs/report/psi_final_version_formatted.docx .

Viscusi, G., Spahiu, B., Maurino, A., & Batini, C. (2014). Compliance with open government data policies: An empirical assessment of Italian local public administrations. *Information Polity: The International Journal of Government & Democracy in the Information Age*, *19*(3/4), 263–275. https://doi.org/10.3233/IP-140338

World Wide Web Foundation (2015*). Open data barometer global report*, third edition. Retrieved on June 19, 2019 from Retrieved from the Open Data Barometer website: http://opendatabarometer.org/doc/3rdEdition/ODB-3rdEdition-GlobalReport.pdf.

Worthy, B. (2015). The Impact of Open Data in the Uk: Complex, Unpredictable, and Political. *Public Administration*, *93*(3), 788–805. https://doi.org/10.1111/padm.12166

User Engagement with Scholarly Twitter Mentions: A Large-scale and Cross-disciplinary Analysis

Zhichao Fang[1]

[1] *z.fang@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (The Netherlands)

Abstract

This study investigates the extent to which scholarly Twitter mentions of scientific papers are engaged with by Twitter users through four types of user engagement behaviors, i.e., liking, retweeting, replying, and quoting. On the basis of a dataset consisting of nearly 8.7 million scholarly Twitter mentions, our results show that there are 55.4% of them have been engaged with by Twitter users through at least one engagement behavior, whereas the remaining did not attract any user engagement. Liking and retweeting are the most common user engagement behaviors. From the disciplinary perspective, scholarly Twitter mentions pertaining to Social Sciences and Humanities are more likely to trigger user engagement. Twitter engagement indicators (i.e., likes, retweets, replies, and quotes) are moderately or strongly correlated with each other, but are weakly correlated with scholarly impact indicators (i.e., citations and Mendeley readers). The analysis of user engagement uncovers the impact of scholarly Twitter mentions on the Twitter universe, shedding light on the characterization of deeper levels of Twitter reception of science information.

Introduction

Twitter mentions of scientific papers (hereinafter "scholarly Twitter mentions") are widespread, thus providing considerable evidence for quantitatively studying science-social media interactions (Costas et al., 2020). The presence of scientific papers on the Twitter platform also offers the possibility for users to engage with science-focused discussion in the Twittersphere (Fang et al., 2021). However, how Twitter users interact with scholarly Twitter mentions, which was specifically referred to as "secondary Twitter metrics" by Díaz-Faes et al., (2019), has not been extensively studied in existing altmetric literature. Therefore, this study aims to present a large-scale and cross-disciplinary analysis of user engagement situations of scholarly Twitter mentions. Specifically, this study seeks to address the following explicit research questions:
RQ1. To what extent are scholarly Twitter mentions engaged with by Twitter users through different engagement behaviors (i.e., liking, retweeting, replying, and quoting)?
RQ2. Do engagement situations of scholarly Twitter mentions vary across subject fields of science?
RQ3. How do Twitter engagement indicators of scholarly Twitter mentions correlate with scholarly impact indicators (i.e., WoS citations and Mendeley readers) of tweeted papers?

Data and methods

Dataset

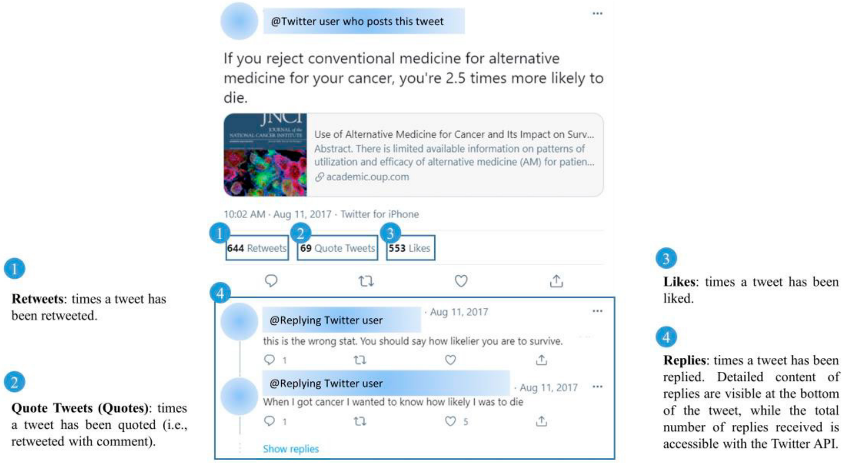
We retrieved a total of 6,229,001 Web of Science-indexed (WoS) papers published between 2016 and 2018 from the CWTS in-house database, and searched their scholarly Twitter mentions recorded by Altmetric.com until October 2019. For the matching with Altmetric.com data, WoS papers are restricted to those with DOI or PubMed ID assigned. On the whole, there are 2,035,648 WoS papers with at least one scholarly Twitter mention received, totally generating 8,692,499 unique scholarly Twitter mentions.[1] Note that to explore user engagement behaviors around scholarly Twitter mentions, in this study the analyzed tweets are limited to those can be engaged with through the engagement functionalities provided by Twitter. Therefore, as listed in Table 1, there are four main types of scholarly Twitter mentions considered in this study, including original tweet, reply tweet, quote tweet, and hybrid tweet. Their proportions in our dataset are also presented in Table 1.

Table 1. The four tweet types with engagement functionalities.

Tweet type	*Concept*	*Number*	*Proportion*
Original tweet	An original tweet is a tweet originally posted by a Twitter user.	7,037,233	81.0%
Reply tweet	A reply tweet is a response to a tweet.	449,544	5.2%
Quote tweet	A quote tweet is a retweet with comment added.	1,174,714	13.5%
Hybrid tweet	A hybrid tweet is a reply tweet with link(s) to other tweet(s) attached. A hybrid tweet is both a reply tweet and a quote tweet.	31,008	0.3%

Twitter engagement indicators

As the tweet example given in Figure 1, at the bottom of a tweet there are four primary engagement indicators displayed and publicly available through the Twitter API, including likes, retweets, replies, and quotes. We focused on these four Twitter engagement indicators and collected them for the 8.7 million scholarly Twitter mentions in our dataset with the Twitter API in February 2021.



@Twitter user who posts this tweet

If you reject conventional medicine for alternative medicine for your cancer, you're 2.5 times more likely to die.

Use of Alternative Medicine for Cancer and Its Impact on Surv...
Abstract. There is limited available information on patterns of utilization and efficacy of alternative medicine (AM) for patien...
academic.oup.com

10:02 AM · Aug 11, 2017 · Twitter for iPhone

1 644 Retweets
2 69 Quote Tweets
3 553 Likes

4

@Replying Twitter user · Aug 11, 2017

this is the wrong stat. You should say how likelier you are to survive.

1

@Replying Twitter user · Aug 11, 2017

When I got cancer I wanted to know how likely I was to die

1 5

Show replies

1 **Retweets**: times a tweet has been retweeted.

2 **Quote Tweets (Quotes)**: times a tweet has been quoted (i.e., retweeted with comment).

3 **Likes**: times a tweet has been liked.

4 **Replies**: times a tweet has been replied. Detailed content of replies are visible at the bottom of the tweet, while the total number of replies received is accessible with the Twitter API.

Figure 1. The four Twitter engagement indicators of a tweet example.

CWTS publication-level classification

To compare the user engagement situations of scholarly Twitter mentions across subject fields of science, we applied the CWTS publication-level classification system (Waltman & Van Eck, 2012) to assign scholarly Twitter mentions with subject field information based on their mentioned scientific papers. The CWTS classification clusters WoS papers into micro-level fields based on their citation relationships. These micro-level fields are then algorithmically assigned to five main subject fields of science, including *Social Sciences and Humanities*

(SSH), *Biomedical and Health Sciences* (BHS), *Physical Sciences and Engineering* (PSE), *Life and Earth Sciences* (LES), and *Mathematics and Computer Science* (MCS).[2] For our dataset, a total of 7,336,299 scholarly Twitter mentions (accounting for 84.4%) refer to scientific papers with the subject field information assigned by the CWTS classification system, involving 3,992 micro-level fields. This set of scholarly Twitter mentions and micro-level fields was drawn as a subsample for studying the subject field differences of user engagement. Figure 2 shows the distribution of the micro-level fields in the subsample.

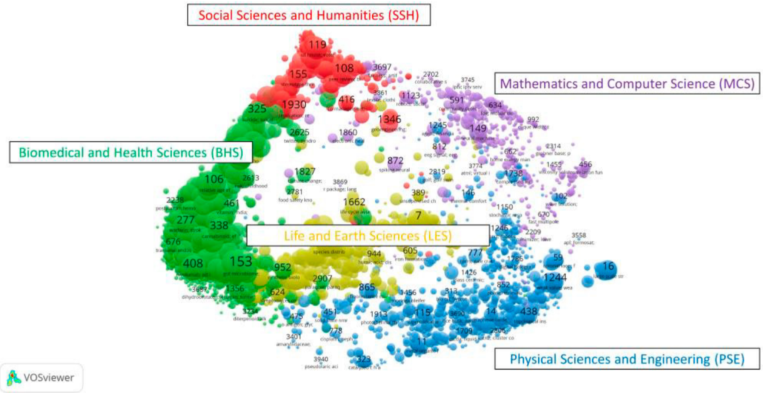


Figure 2. Distribution of micro-level fields of the five subject fields of science. Each node represents a micro-level field with the size determined by the number of scholarly Twitter mentions inclusive. Micro-level fields belonging to the same subject field are indicated in the same color.

Analytic approaches

To unravel the overall user engagement situations of scholarly Twitter mentions, descriptive statistics of Twitter engagement indicators are first conducted. Besides, we compared the relationships between the analyzed Twitter engagement indicators and scholarly impact indicators. To this end, WoS citations and Mendeley readers were selected as indicators reflecting the scholarly impact of papers. The former was retrieved from the CWTS in-house WoS database recording citations accumulated up to March 2020, and the latter was collected through the Mendeley API in July 2020. We aggregated both scholarly impact indicators and Twitter engagement indicators at the paper level and then performed Spearman correlation analyses amongst them.

Results and discussion

Overall user engagement around scholarly Twitter mentions

Among the 8.7 million scholarly Twitter mentions, 55.4% have been engaged with through at least one of the four analyzed engagement behaviors, namely, the overall coverage of Twitter engagement among scholarly Twitter mentions is 55.4%. More specifically, Table 2 presents the descriptive statistics of the four engagement indicators to reflect the extent to which scholarly Twitter mentions are engaged with through different behaviors. Liking is the most common engagement behavior, with 47.7% of scholarly Twitter mentions being liked by Twitter users. Followed by retweeting which involves 35.7% of scholarly Twitter mentions. In contrast, replying and quoting, the two engagement behaviors with original comment added by Twitter users, are relatively scarce, with only 10.3% of scholarly Twitter mentions being replied and 8.4 % being quoted at least once.

Table 2. Descriptive statistics of the four engagement indicators.

Engagement	*Total number*	*Coverage*	*Mean*	*Min*	*Q1*	*Q2*	*Q3*	*Max*	*SD*
Likes	26,071,539	47.7%	3.00	0	0	0	2	28,368	24.23
Retweets	15,669,405	35.7%	1.80	0	0	0	1	33,009	21.15
Replies	1,388,947	10.3%	0.16	0	0	0	0	1,033	1.24
Quotes	1,347,088	8.4%	0.15	0	0	0	0	804	1.47

Users engagement around scholarly Twitter mentions across subject fields

For each type of user engagement behavior, the micro-level fields in different subject fields of science are plotted in Figure 3 based on the coverage of corresponding engagement among their inclusive scholarly Twitter mentions.

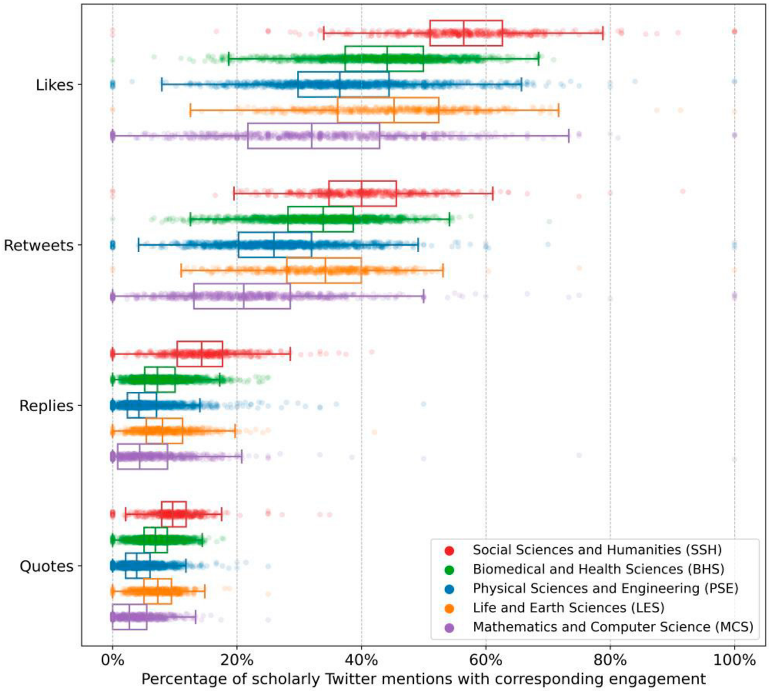


Figure 3. Coverage of the four types of user engagement among scholarly Twitter mentions across the five subject fields of science.

Generally speaking, the disciplinary differences are similar across the four Twitter engagement indicators. Social Sciences and Humanities is the subject field with related scholarly Twitter mentions more likely to be engaged with by Twitter users, regardless of what kind of user engagement behavior. Next, scholarly Twitter mentions in the fields of both Life and Earth Sciences and Biomedical and Health Sciences attract relatively high levels of user engagement as well. By comparison, there are less scholarly Twitter mentions in relation to Physical Sciences and Engineering and Mathematics and Computer Science being engaged with on Twitter. Similar to the more active Twitter mention events observed for social sciences, health sciences, and life sciences-related research outputs (Costas et al., 2015; Haustein et al., 2015), there are more active user engagement behaviors around scholarly Twitter mentions of these subject fields, suggesting a consistent Twitter users' interest in sharing, discussing and disseminating research progress in terms of society, health and environmental problems.

Correlations between scholarly impact and Twitter engagement

We performed Spearman correlation analyses among the two scholarly impact indicators (i.e., WoS citation and Mendeley readers) and the four Twitter engagement indicators, and Figure 4 shows the results. As confirmed by previous studies (Zahedi et al., 2014), the two scholarly impact indicators, WoS citations and Mendeley readers, keep a moderate correlation. However, the correlations between scholarly impact indicators and Twitter engagement indicators are weak, reinforcing the idea that scholarly metrics and Twitter metrics reflect different aspects of the impact that scientific papers made. Twitter engagement indicators are moderately or strongly correlated with each other, indicating intrinsic relationships among user engagement behaviors within the Twitter universe.

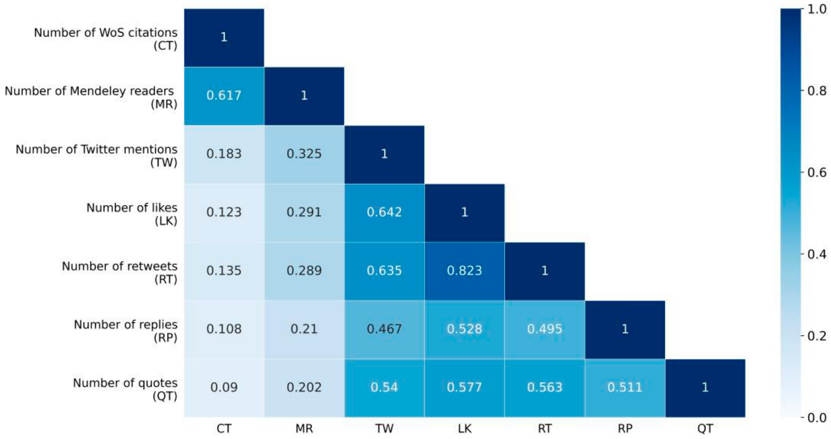


Figure 4. Spearman correlations among scholarly impact indicators and Twitter engagement indicators.

Conclusions and future outlook

This paper presents a large-scale and cross-disciplinary analysis of Twitter user engagement for nearly 8.7 million scholarly Twitter mentions. We found that even though scholarly Twitter mentions widely brought research outputs to the Twitter landscape, only 55.4% of them have been engaged with by Twitter users through at least one of the four engagement behaviors (i.e., liking, retweeting, replying, and quoting), showing a deeper level of Twitter reception of science information. Comparatively speaking, scholarly Twitter mentions about social sciences, life sciences and health sciences exhibit a higher coverage of user engagement. Liking and retweeting, as the two basic user engagement behaviors without additional original commentary added, are the most common and are strongly correlated with each other. In addition to likes and retweets, other Twitter engagement indicators are also moderately correlated, reflecting intrinsic relationships among Twitter users' engagement behaviors. However, the correlation between scholarly impact indicators and Twitter engagement indicators are generally weak. These two types of indicators might tell different stories of the reception and dissemination of scientific knowledge.

For future research, we will take into consideration the features of scholarly Twitter mentions (e.g., the inclusion of hashtags or mentioned users, the originality of tweet texts) as well as the demographics of Twitter users (e.g., number of followers, Twitter activity) to explore what kinds of scholarly Twitter mentions are more likely to trigger user engagement behaviors, thus paving the way toward a better understanding of how and why Twitter users interact with science.

References

Costas, R., Rijcke, S. & Marres, N. (2020). "Heterogeneous couplings": Operationalizing network perspectives to study science-society interactions through social media metrics. *Journal of the Association for Information Science and Technology*, advanced online publication. https://doi.org/10.1002/asi.24427

Costas, R., Zahedi, Z. & Wouters, P. (2015). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10), 2003–2019. https://doi.org/10.1002/asi.23309

Díaz-Faes, A. A., Bowman, T. D. & Costas, R. (2019). Towards a second generation of 'social media metrics': Characterizing Twitter communities of attention around science. *PLoS ONE*, 14(5), e0216408. https://doi.org/10.1371/journal.pone.0216408

Fang, Z., Costas, R., Tian, W., Wang, X. & Wouters, P. (2021). How is science clicked on Twitter? Click metrics for Bitly short links to scientific publications. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24458

Fang, Z., Dudek, J. & Costas, R. (2020). The stability of Twitter metrics: A study on unavailable Twitter mentions of scientific publications. *Journal of the Association for Information Science and Technology*, 71(12), 1455–1469. https://doi.org/10.1002/asi.24344

Haustein, S., Costas, R. & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLoS ONE*, 10(3), e0120495. https://doi.org/10.1371/journal.pone.0120495

Waltman, L. & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378–2392. https://doi.org/10.1002/asi.22748

Zahedi, Z., Costas, R, & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics*, 101(2), 1491–1513. https://doi.org/10.1007/s11192-014-1264-0

[1] Unavailable scholarly Twitter mentions caused by deletion of tweets, or suspension and protection of Twitter users' accounts (Fang et al., 2020) are not included in the dataset.

[2] See more introduction to the CWTS classification system (also known as the Leiden Ranking classification) at: https://www.leidenranking.com/information/fields.

An Alternative Analysis Method for Measuring the Impact of Academic Papers Shared on Social Media and the Number of Citations Obtained based on a Support Vector Machine Algorithm

Adian Fatchur Rochim[1], Faisal Rizky Rahadian[2] and Dania Eridani[3]

[1]adian@ce.undip.ac.id, [2]faisalrr@students.undip.ac.id, [3]dania@ce.undip.ac.id
Universitas Diponegoro, Faculty of Engineering, Dept. of Computer Engineering, Tembalang 50275, Semarang (Indonesia)

Abstract

Information technology affects most aspects of human life. Social Media (MedSos) is an information technology product. One of the uses of social media in the academic world is Altmetrics. This indicator is used to measure the impact or influence of social media on indexed papers. Indonesia is one of the countries with the highest social media users in the world. Therefore, this study is proposed to measure the possibility of a correlation between comments / mentions on papers shared on social media and the number of citations obtained. In solving this problem, we propose a method that uses Text Mining to perform Natural Language Processing (NLP) so that machines can understand the meaning of human language and maximize class distance; we used the Support Vector Machine (SVM) algorithm method for classifying opinions in a scientific article. We found that publications shared on social media will have more citations. Papers that have a greater number of positive sentiments will have a large number of citations, whereas the number of tweets on a paper has no effect on the value of positive sentiments and tends to be more contradictory.

Introduction

In its development, information technology is very influential in almost all aspects of life. One example of the results of information technology is social media. Social media is used by people who use it to meet needs, support activities, and open up opportunities to realize new hopes (Akram et al., 2017). The research progress creates this and is one of the developments in communication technology.

Currently, social media has grown rapidly along with technological advances and has penetrated various layers and groups of society. However, there are still few matters concerning social media's impact or influence on researchers' indexed papers in its development. Therefore, sentiment analysis is required regarding this matter based on several factors and criteria. To obtain data with positive or negative review comments that affect the study's h-index or number of citations. Although the h-index has its drawbacks, it is still used as long as there is no better substitute (Rochim et al., 2020).

In previous research, Xiaoli in 2020 discussed dynamics of topic inheritance research and topic innovation by using cross-collection topic models and measuring direct and indirect scientific influence through "citations" (Chen and Han, 2020). Previous research entitled "Sentiment Analysis using a Support Vector Machine" (Nomleni, 2015) discussed the classification of textual documents into several classes, such as positive and negative sentiments and the effects and benefits of sentiment analysis. In this study, the classification of public complaints against the government on social media, Facebook and Twitter, was used with Indonesian language data using a Support Vector Machine (SVM) run in a distributed computer using Hadoop. A study entitled "Adapting SVM for Natural Language Learning: Case Studies Involving Information Extraction" (Li et al., 2008) discussed two techniques to help SVM with the NLP problem's two unique features: unbalanced training data and difficulty obtaining adequate training data. The research problem is how to measure the impact of papers in social media that correlate with several citations. Jason in 2010 stated that Altmetrics (social media in scholars) would become scientometrics 2.0 (Jason et al., 2010). Most of all database indexers, i.e., Scopus, IEEE Xplore used Altmetrics to figure impact profiles of authors from social media.

Methodology

The methodology used in the research adopts an experimental method from various international papers obtained from the Altmetrics website. It discusses how social media's positive comments come to the number of citations in indexed scientific papers made using the Support Vector Machine method to produce some descriptive analysis.

This research begins with data collection totaling 50 different scientific papers from the Altmetrics website (accessed April 13, 2020). This study's data type is primary data as test data obtained from the Altmetrics website and secondary data obtained from Kaggle created by Ali Toosi as training data. The dataset contains 50 datasets of scientific papers, with 700 comments for each dataset. After all the data are collected, the data translation process into English will then be carried out.

Sentiment Analysis Process

Figure 1 illustrates stages of data or documents that enter the system, which are then carried out, cleaning the document to eliminate unnecessary words. After that, the parsing or tokenization process is carried out to divide or break the document into terms based on the stop word and then delete it to filter words or documents. Finally, the stemming process is carried out to obtain common words according to applicable standards.

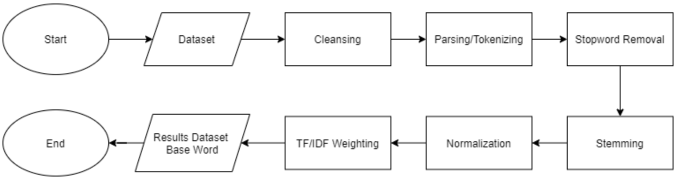


Figure 1. Flow diagram sentiment analysis process

1) Cleansing used aims to remove unnecessary words and characters as noise reduction; Perform cleaning to change all capital letters and documents to lowercase and remove characters other than punctuation, repeating letters, and hyperlinks. Table 1 (Process no.1) presents samples of the cleansing process.

Table 1. Sample of Processing.

No.	Process	Sentence	Results
1.	Cleansing	~~RT @ learn_learning3:~~ It requires large amounts of the image of the person who is the generation source of the video~~.~~	It requires large amounts of the image of the person who is the generation source of the video.
2.	Parsing and Tokenization	but it requires large amounts of the image of the person who is the generation source of the video.	[but, it, requires, large, amounts, of, the, image, of, the, person, who, is, the, generation, source, of, the, video]
3.	Stopword	[~~but, it~~requires, large, amounts, ~~of, the,~~ image, ~~of, the,~~ person, ~~who, is, the,~~ generation, source~~, of, the,~~ video]	[requires, large, amounts, image, person, generation, source, video]
4.	Stemming	[requires, large, **amounts**, image, person, **generation**, source, video]	[require, large, **amount**, image, person, **generate**, source, video]

2) Parsing and tokenization were used to divide large parts of the document into words chop off each word in the text and change all uppercase letters to lowercase. Table 1 (Process no.2) presents samples of the parsing and tokenization process that divides or breaks a large part of a document (sentence) into words; the process chops off each word in the text.

3) Stopword Removal was used to filter words or documents identified as conjunctions, articles, and prepositions. Stopword Removal is the process of removing unnecessary words based on the stopword dictionary in English ('between,' 'yourself,' 'but,' 'again,' 'there,' and so forth). Such words have no meaning. Table 1 (Process no.3) presents the process of stopword removal.

4) Stemming is used to find a root word in a word and remove prefixes, suffixes, and combinations of prefixes and suffixes (Zainuddin et al., 2014). Table 1 (Process no.4) presents samples of stemming that changes the processing of words into basic words by eliminating prefixes, insertions (infixes), suffixes, and combinations of prefixes and suffixes.

5) Normalization is used to measure variable values on a broader scale and change the previous value to become 1. The data should be scaled before training is carried out on the data normalized with mean = 0 and standard deviation = 1. The normalization formula is

$$\text{New Value} = \frac{[Old\ Value(Average)]}{Standard\ Deviation} \tag{1}$$

Data that have been normalized will be divided into training data and testing data using the cross-validation method.

6) Weighting. In weighting this term, the results of the stemming process will be used in calculating the number of documents' Term Frequency (TF), the number of documents that have a term (DF), and the IDF value such as the formula (Nomleni, 2015), (Amrizal et al., 2018). Table 2 presents a sample of the weighting process.

Table 2. Weighting.

Tweet	*Rank*
Ability	0.50
Able	0.69
Work	1.00
Year	0.71

7) TF-IDF Weighting. The method used to find a representation of the value of the data set is trained, and the results form a vector between a document and a word. This method combines a two-weight calculation concept, specifically Term Frequency (TF) and Inverse Document Frequency (IDF). TF determines the word-for-word weight in a document and IDF serves to construct contributions from words into a document. Term Frequency is the frequency of occurrence of words (F) in a sentence (D), and Document Frequency (DF) is the number of sentences in which a word (F) appears (Nomleni, 2015). This word weighting will produce a word weight value, which indicates the importance of each word in the document (Vijayarani et al., 2015). This TF-IDF weighting calculation is formulated in the following equation:

$$\text{idf} = \log\left(\frac{N}{df}\right), \tag{2}$$

where

N = number of document collections, and df = number of documents containing pre-determined term (f)

$$\mathrm{W_{dt}} = tfdt \times idft, \tag{3}$$

where

W_{dt} = weight term (t) against the document, (d)

idf = inversed document frequency (log/(N/df)).

$tfdt$ = number of occurrences of term (t) in the document (d), and

Support Vector Machine Algorithm (SVM)

SVM is used as a classification algorithm to find the best hyperplane by maximizing the distance between classes. Classification is done to find a hyperplane or the boundary line (Decision Boundary), which separates a class from another class.

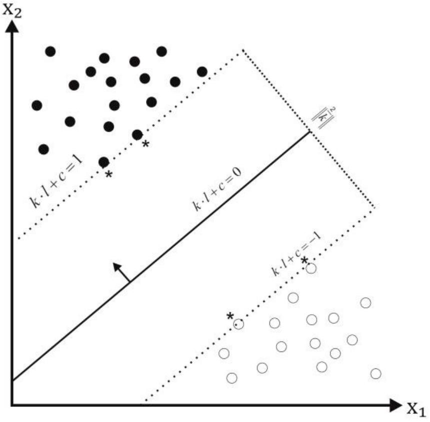


Figure 2. When the wrong hyperplane is positive (+1) and negative (-1).

Figure 2 shows a solid line that is the best hyperplane found by the SVM algorithm, located right in the middle of the second class. The distance between the hyperplane and the data object is different from the near (outermost) class, given an asterisk or star and the data object. The outermost closest to the hyperplane is called the Support Vector (Marafino et al., 2014).

Precision, Recall, and F-Measure

After the SVM classification process is complete, testing of the classification is carried out to measure the performance value of the system that has been created. The formulas used for Precision, Recall, Accuracy, and F-Measure are as follows:

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{4}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{5}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \text{and} \tag{6}$$

$$\text{F} - \text{Measure} = 2 * \frac{precision*recall}{precision+recall}. \tag{7}$$

Precision is the ratio of true positive predictions compared to total positive predicted results (Irawan et al., 2018), Recall is the ratio of true positive predictions to total true positive data (Flach et al., 2015). Accuracy is the number of correctly predicted documents divided by the total number of documents, and F-Measure is a weighted average comparison of precision and recall (Lipton et al., 2014).

Results and Discussion

From the results of the research that has been conducted, the results obtained from the experiment are Precision, Recall, F-Measure, Accurate, the number of citations, and positive and negative sentiments from each paper.

Table 2 presents the confusion matrix and results of SVM. This method is used to avoid a perfect score but fails to predict anything in the not-yet-visible data (overfitting).

Table 2. Confusion matrix and results of SVM.

Actual Data	*Predict Data*		*Precision*	*Recall*	*F-Measure*	*Accuracy*
	Positive	*Negative*				
Positive	19	6	0.76	1	0.86	0.96
Negative	0	115	0.95	1	0.97	

Positive class:

$$Precision = \frac{TP}{TP+FP} = \frac{19}{19+6} = 0.76 \qquad F-Measure = 2 \times \frac{precision \times \ recall}{precision + recall} = 2 * \frac{0.76 \times \ 1}{0.76+1}$$

$$Recall = \frac{TP}{TP+FN} = \frac{19}{19+0} = 1 \qquad = 0.86$$

Negative class:

$$Precision = \frac{TN}{TN+FP} = \frac{115}{115+6} = 0.95$$

$$Recall = \frac{TN}{TN+FN} = \frac{115}{115+0} = 1 \qquad F-Measure = 2 \times \frac{precision * recall}{precision + recall} = 2 \times \frac{0.95 * 1}{0.95+1} = 0.97$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{19+115}{19+115+6+0} = \ 0.957 \times \ 100\% = 96\%$$

Test Results

The measurement results for each paper obtained by the average Precision, Recall, and F-Measure using the SVM classification are presented in Table 3.

Table 3. Sentiment analysis results.

Paper id	*Precision*	*Recall*	*F-Measure*	*Accuracy*	*Number of citations*	*P%*	*N%*
1	0.91	0.96	0.93	0.94	881	70.1	29.9
2	0.99	0.94	0.96	0.99	1875	91.6	8.4
3	0.99	0.94	0.96	0.99	354	85.3	14.7
4	0.99	0.87	0.92	0.99	1155	93.8	6.2
5	0.98	0.95	0.96	0.97	3190	78.7	21.3
6	0.99	0.93	0.95	0.99	641	77.1	22.9
7	0.97	0.94	0.95	0.97	799	51.7	48.3
8	0.99	0.93	0.95	0.99	589	90.3	9.3
9	0.97	0.66	0.74	0.95	1502	91.3	8.7
10	0.98	0.83	0.89	0.97	408	86.5	13.5
11	0.96	0.85	0.89	0.94	2721	79.8	20.2
12	0.97	0.80	0.86	0.94	2981	89.1	10.9
13	0.94	0.82	0.86	0.91	3041	76.4	23.6
14	0.93	0.66	0.71	0.87	415	78.4	21.6
15	0.94	0.82	0.86	0.91	701	76.4	23.6
16	0.91	0.75	0.78	0.85	454	76.3	23.7

Paper id	*Precision*	*Recall*	*F-Measure*	*Accuracy*	*Number of citations*	*P%*	*N%*
17	0.90	0.62	0.65	0.82	265	73.9	26.1
18	0.98	0.85	0.90	0.97	4709	90.2	9.8
19	0.98	0.94	0.96	0.98	30	86.2	13.8
20	0.95	0.95	0.95	0.97	1316	78.7	21.3
21	0.97	0.81	0.87	0.94	312	87.2	12.8
22	0.89	0.75	0.77	0.82	485	72.9	27.1
23	0.96	0.86	0.90	0.94	510	79.8	20.2
24	1.00	1.00	1.00	1.00	282	53.1	46.9
25	0.955	0.78	0.83	0.92	403	75.6	26.4
26	0.955	0.725	0.78	0.91	287	80.3	19.7
27	1.00	1.00	1.00	1.00	197	90.9	9.1
28	0.48	0.50	0.49	0.97	516	88.7	11.3
29	0.97	0.85	0.90	0.96	1145	82.5	17.5
30	0.98	0.85	0.90	0.97	312	89.8	10.2
31	0.48	0.50	0.49	0.97	730	95.2	4.8
32	0.98	0.80	0.86	0.97	954	91.4	8.6
33	0.48	0.50	0.49	0.97	151	96.8	3.2
34	0.92	0.66	0.71	0.86	1242	76.8	23.2
35	0.95	0.94	0.95	0.95	510	68.4	31.6
36	0.87	0.86	0.86	0.89	1588	75.0	25.0
37	0.99	0.87	0.92	0.98	906	91.5	8.5
38	0.98	0.81	0.87	0.96	677	88.1	11.9
39	0.96	0.89	0.92	0.94	729	66.6	33.4
40	0.98	0.91	0.94	0.97	21	71.7	28.3
41	0.95	0.87	0.90	0.92	88	64.6	35.4
42	0.97	0.83	0.88	0.95	228	83.1	16.9
43	1.00	1.00	1.00	1.00	235	91.6	8.4
44	0.91	0.89	0.89	0.89	32	63.8	36.2
45	0.89	0.61	0.62	0.79	608	82.8	17.2
46	0.93	0.70	0.75	0.88	757	74.4	25.6
47	0.98	0.75	0.82	0.97	187	89.6	10.4
48	0.90	0.82	0.85	0.93	409	82.4	17.6
49	0.93	0.65	0.69	0.87	1708	88.9	11.1
50	0.97	0.75	0.82	0.96	440	90.2	9.8
Score	0.93	0.82	0.85	0.94	893.52	81.1	18.9

Figure 3 illustrates the comparison of the correlation between the total number of tweets and positive sentiment. In measuring validity between citations data and positive data, we used Spearman's Rank-Order Correlation method. We found that the citations index has a reasonably low correlation. The correlation between citations and positive sentiment is 0.085. Even though this number is very small, it indicates that positive sentiment influences citations. It can be concluded that the higher the positive sentiment, the larger the number of citations. Figure 4 presents a graph between the number of tweets and the sentiment in the opposite direction. Furthermore, using Spearman's Rank-Order Correlation test between the number of tweets and the positive sentiment gave a negative value of −0.183. However, it can be concluded that it does not affect the sentiment score.

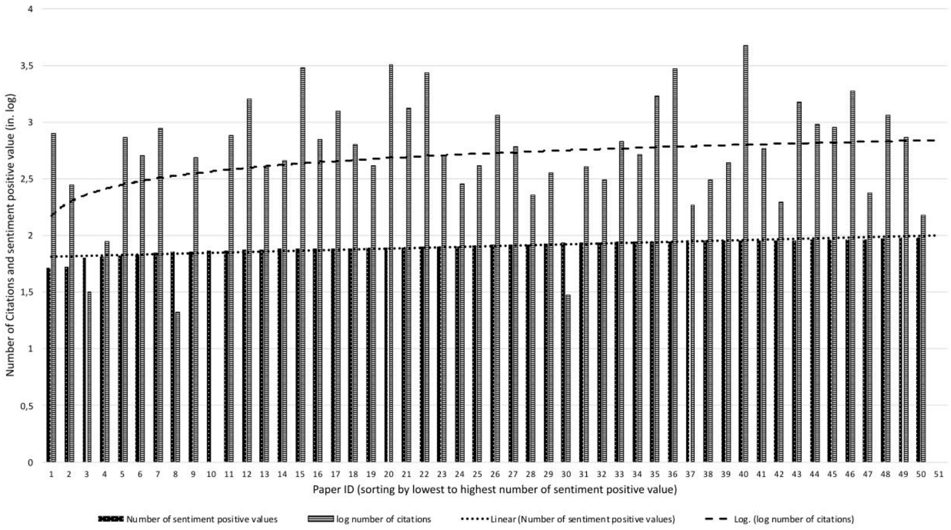


Figure 3. Distribution of the correlation between positive sentiment and number of citations.

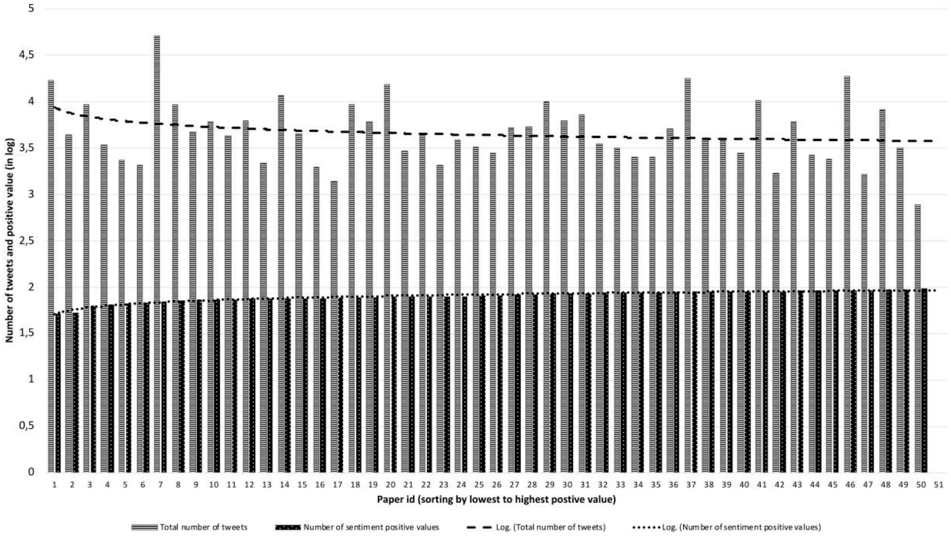


Figure 4. Distribution of the correlation between total tweets and positive sentiment.

Conclusion

We have found that there are correlations among the number of citations of papers obtained, the number of tweets of papers, and number of positive sentiments on papers. We have found that there is a correlation of 0.08 between the number of citations obtained and the number of positive sentiments on a paper. The correlation test between the number of tweets and the number of positive sentiments is −0.183. We can conclude that the number of positive sentiments obtained by the paper will affect the number of citations it gets. The number of tweets obtained by papers on social media has no impact on the number of citations they obtain.

However, this influence is not so significant that it needs to be investigated more. It indicates that the papers that have more several positive sentiments have a larger number of citations.

Acknowledgment

This research was financially supported by the Ministry of Research and Technology, Republic of Indonesia, through Fundamental Research Grant No. B/112/E3/RA.00/2021.

References

Akram, W. & Kumar, R. (2017). A Study on Positive and Negative Effects of Social Media on Society. *International Journal of Computer Sciences and Engineering*, 5 (10), 351–54.

Amrizal, V. (2018). Application of the Term Frequency Inverse Document Frequency (Tf − Idf) and Cosine Similarity Methods in the Information Retrieval System to Determine Web-Based Hadiths (Case Study: Hadith Sahih Bukhari − Muslim*). Journal Teknik Informatika*, 11(2), 149−164, doi:10.15408/jti.v11i2.8623.

Chen, X. & Han, T. (2020). A Micro Perspective of Research Dynamics Through 'Citations of Citations' Topic Analysis. *Journal of Data and Information Science*, 5(4), 1–16.

Flach, P. A. & Kull, M. (2015). Precision−Recall−Gain Curves: PR Analysis Done Right. *Neural Information Processing Systems*, 28, 838−846.

Irawan, F. & Samopa. F. (2018). A Comparative Assessment of Random Forest and SVM Algorithms Using Combination of Principal Component Analysis and SMOTE for Accounts A Comparative Assessment of Random Forest and SVM Algorithms, Using Combination of Principal Component Analysis and SM. *The 2nd International Seminar of Contemporary Research on Business & Management 2018*.

Priem, J. & Hemminger. B. (2010). Scientometrics 2.0: New Metrics of Scholarly Impact on the Social Web, *First Monday*, 15 (7).

Li, Y., Bontcheva, K., & Cunningham. H. (2008). Adapting SVM for Natural Language Learning: A Case Study Involving Information Extraction. *Natural Language Engineering*, 15(2),1–25.

Lipton, Z. C., Elkan, C. & Naryanaswamy, B. (2014). Optimal Thresholding of Classifiers to Maximize F1 Measure. *Lecture Notes in Computer Science,* 8725 (2), 225–239, doi:10.1007/978-3-662-44851-9_15.

Manek, A.S., Shenoy, P.D., Mohan, M.C. & Venugolan, K.R. (2017). Aspect Term Extraction for Sentiment Analysis in Large Movie Reviews Using Gini Index Feature Selection Method and SVM Classifier. *World Wide Web*, 20(2), 135–154, doi:10.1007/s11280-015-0381-x.

Marafino, B. J., Davies, J.M., Bardach, M.S., Dean, M.L. & Dudley, R.A. (2014) N-Gram Support Vector Machines for Scalable Procedure and Diagnosis Classification, with Applications to Clinical Free Text Data from the Intensive Care Unit. *Journal of the American Medical Informatics Association*, 21(5), 871–875, doi:10.1136/amiajnl-2014-002694.

Nomleni, P. (2015). Sentiment Analysis using Support Vector Machine. *Thesis, Institute of Technology Sepuluh Nopember*, 2015, pp. 5–6.

Rochim, A. F, Muis, F.A. & Sari, R.F. (2020). A Discrimination Index Based on Jain's Fairness Index to Differentiate Researchers with Identical H-index Values. *Journal of Data and Information Science*, 5(4), 5-18.

Vijayarani, S. & Gurusamy, V. (2015) Preprocessing Techniques for Text Mining: An Overview. *International Journal of Computer Science and Communication Networks*, 5(1), 7–16.

Zainuddin, N. & Selamat, A. (2014). Sentiment Analysis Using Support Vector Machine. *I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology*, pp. 333–337, doi:10.1109/I4CT.2014.6914200.

Doing as the Romans do? A comparative analysis of the thematic profiles of newcomers and already existing faculty at universities

Márcia R. Ferreira[1] and Rodrigo Costas[2]

[1]*ferreira@csh.ac.at*
Complexity Science Hub Vienna, Vienna (Austria)
Vienna University of Technology (TU Wien), Vienna (Austria)

[2]*rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Leiden University, Leiden (the Netherlands)
DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy,
Stellenbosch University, Stellenbosch (South Africa)

Abstract

This paper analyses the relationship between academic mobility and the structure of universities' thematic portfolios. To do this, we examined millions of authorships on a global scale through the study of authors' affiliation trajectories across institutions and scientific topics. Specifically, we analyzed all articles and reviews available in the Dimensions database and used a publications-based classification of 4,064 scientific topics to map the publishing activity of 4,078,913 disambiguated authors. Examining a sample of 1,189 Leiden Ranking universities over the recent period of 2015-2020, we show that after researchers move to a new institution, they tend to publish on similar research topics to those of their peers already at the same institution. This result suggests that mobile researchers' choices of institution may be driven either by potential thematic affinity (*homophily*) or by the adoption of the topics already developed at the new institutions (*adaptation*). Areas of future development are also suggested.

Introduction

Mobile researchers play an important role in the process of discovery. They drive knowledge exchanges (OECD, 2010) that are vital for the transfer of "tacit knowledge", which cannot always be transmitted through formal communication channels (Gertler, 2003). When researchers move, they bring with them different perspectives and ideas. These perspectives are considered to be important for knowledge recombination (Ganguli, 2015; Stephan & Levin, 2001), which leads to new opportunities for scientific innovation (Franzoni, Scellato & Stephan, 2018). It is therefore important to gain a better understanding of the skills and expertise that mobile researchers bring to host institutions, as well as how this can influence their own and their institutions' ability to explore and develop uncharted scientific areas.

Despite the importance of mobility flows to scientific innovation, much of how institutions attract new scholars, and how prospective newcomers compare to the native faculty in terms of the topics in which they publish remains unclear. This paper attempts to address this challenge by comparing the topics of publications by mobile researchers to the knowledge base of receiving institutions based on the topics of native researchers. The question we address is *how do the research areas of newcomers (researchers moving to a new university) and native researchers (researchers who begin their publishing careers at a university) differ within and across universities?*

Database and methods

We use the Centre for Science and Technology Studies (CWTS) in-house version of the Dimensions database. This database contains publication data spanning several decades. This analysis focuses on the universities featured in the most recent release of the Leiden Ranking (https://www.leidenranking.com/). This approach allows us to track mobility between a homogeneous set of institutions (Macháček et al, 2020). While we present the analysis of the restricted set of universities, our indicators capture the mobility events of researchers with institutions not covered by the Leiden Ranking[1].

Classification of authorships into 'newcomers' and 'natives'

We start by identifying the first affiliation of researchers based on their first publication in the whole Dimensions database (i.e., the *original* affiliation of the author) according to the time sequence of their overall publication history. This means that we look at scholars' entire publication history and not just at their publication activity in the period under analysis (i.e., 2015-2020). Then, as a second step, we identify all the authorships of each researcher at her original affiliation or *native authorships* and identify all the authorships of each researcher as a *newcomer* to another university[2] or *newcomer authorships*. We aggregate the two kinds of authorship at the institutional level – in this case for Leiden Ranking universities – in the period between 2015 and 2020. Our choice for a recent period is because it allows us to have a more contemporary overview of the workforce composition of institutions' in terms of mobility. We also consider the micro-fields of the publications of the authorships previously identified and split the set of overall authorships of a university into two:

1. *Native authorships*. Micro-field information weighted by the number of authorships by those researchers whose first affiliation was already at the current institution in the period. This can be interpreted as the micro-fields profile vector of university natives in a given period.
2. *Newcomer authorships*. Micro-field information weighted by the number of authorships by those researchers whose first affiliation was not at their current institution in the period under analysis. This can be interpreted as the micro-fields profile vector of university newcomers in a given period.

Each university is characterised by these two micro-fields vectors. This is used to compare the overall research profile (dis)similarity, as captured by micro-fields, between institutional natives and newcomers. To compute the similarity between the vectors we use the cosine similarity measure as described in the next section.

Cosine Similarity

Starting from the overall set of authorships in the output of a university in the period 2015-2020, we measure the cosine similarity of the micro-fields from the newcomers to a university, and we compare it with those of the native researchers.

As micro-fields we use the publication-level classification develop by Waltman & van Eck (2012). This publication-level classification is composed by 4,064 scientific micro-fields obtained algorithmically based on citation relationships among publications. This fine-grained classification allows for the advanced study of the thematic profile of universities and individual researchers from a topical point of view, in which micro-fields can be seen as specialized research areas.

Similarity is constructed using the standard cosine similarity measure to compare the vectors of research topics of newcomer authorships (vector **A**) and the vector of research topics of native authorships (vector **B**). Consider for example, a university with a single newcomer. The newcomer author $\boldsymbol{a}$ to university $\boldsymbol{u}$ in the period of the analysis, wrote three papers assigned to three topics $\boldsymbol{i}$. Let us assume that one micro-field, say $\boldsymbol{\alpha}$, has been used in all three papers. The

[1] For example, mobility linkages of researchers with other local universities or other types of research organizations (e.g., umbrella research organizations like Inserm, CSIC, CNRS, etc.) are also considered in the identification of mobile researchers.

[2] It is possible that a publication contains both newcomer authorships and native authorships if the same publication is authored both by natives and newcomers. In these cases, the publication is weighted by the number of authorships that are '1' (native) and '0' (newcomer).

component[3] $\boldsymbol{\alpha}$ will be 3. Thus, the components quantify the number of publications of an author in each micro-field.
More formally, we calculate the cosine similarity as follows

$$\theta = \cos(\gamma) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|A\|_2 \|B\|_2}$$

, is defined for each pair of vectors $\boldsymbol{A}$ and $\boldsymbol{B}$.
Other indicators, such as the number of researchers, the number of micro-fields, the number of publications (**P**), the proportion of natives and the proportion of newcomers, are also provided (See Table 1).

Results

Table 1 provides some summary statistics of the variables used in the analysis.

Table 1. Descriptive Statistics.

Statistic	*Number Researchers*	*Number Micro-Fields*	*P*	*Proportion - Natives*[4]	*Proportion - Newcomers*	*Cosine Similarity*
N	1189	1189	1189	1189	1189	1189
Mean	4613.84	1188.23	8703.16	0.65	0.34	0.78
SD	4684.81	558.30	8775.40	0.13	0.13	0.15
Min	2	1	1	0	0	0
25%	1789	778	3368	0.57	0.24	0.73
50%	3043	1079	5771	0.67	0.33	0.82
75%	5483	1542	10443	0.76	0.43	0.89
Max	43016	2984	65289	1	1	1

Institutional cosine similarity distribution

In Fig.1 we compare the distribution of $\boldsymbol{\theta}$ from different perspectives. Fig.1 a) shows that there is a tendency towards a high cosine similarity between the thematic profiles of newcomers and natives across universities, represented by most institutions having $\boldsymbol{\theta} > 0.7$. Thus, institutions were more likely to attract newcomer researchers working in the same topics as the natives. To better understand this result; in Fig.1 b) we also compare the distribution of $\boldsymbol{\theta}$ with a null model obtained considering a randomised assignment of newcomers[5]. The plot clearly supports a legitimate tendency of universities to have newcomers covering similar topics as those developed by native researchers (high values of $\boldsymbol{\theta}$), as the actual distribution (the blue bar) contrasts with the randomized distribution (the red bar). The randomized set is strongly shifted to the left, meaning that when the authorships are randomly assigned to universities, including the corresponding micro-fields of each authorship and the cosine similarity is recalculated, the

[3] Component here refers to the elements of the vectors of authorships.

[4] Note that the proportion of natives and newcomers in the table refers to the distinct number of researchers falling in those categories from the perspective of institutions for the period 2015-2020.

[5] These randomized sets are obtained considering the period 2015-2020 and assigning researchers randomly to universities. The hypothesis is that institutions have a general tendency to pick researchers that are similar to their existing researchers, so if you shuffle the researchers and attribute them randomly to universities, we can check this specific hypothesis.

differences between the authorships of the newcomers and the authorships of the natives are substantially different. In Fig.1 c) we repeat this same analysis but considering the Jaccard index. This is a test of robustness of the results to avoid possible spurious effects induced by sparse vectors in the cosine similarity. The plot confirms the pattern emerging from Fig.1 a). This reassures us that the $\boldsymbol{\theta}$ is not sensitive to spurious effects and high values of $\boldsymbol{i}$. These results demonstrate that universities' newcomers and natives tend to research similar topics. However, as we measure the similarity between newcomers' authorships after they become affiliated with the new university - and not before they join the institution, we cannot discern yet whether newcomers do *adapt* to the thematic profile of their new universities, or whether the newcomer's choice of a new institution is driven by *homophily*, i.e., moving to universities already working on the same topics as the researcher was already working. This is clearly a topic to study in future research.

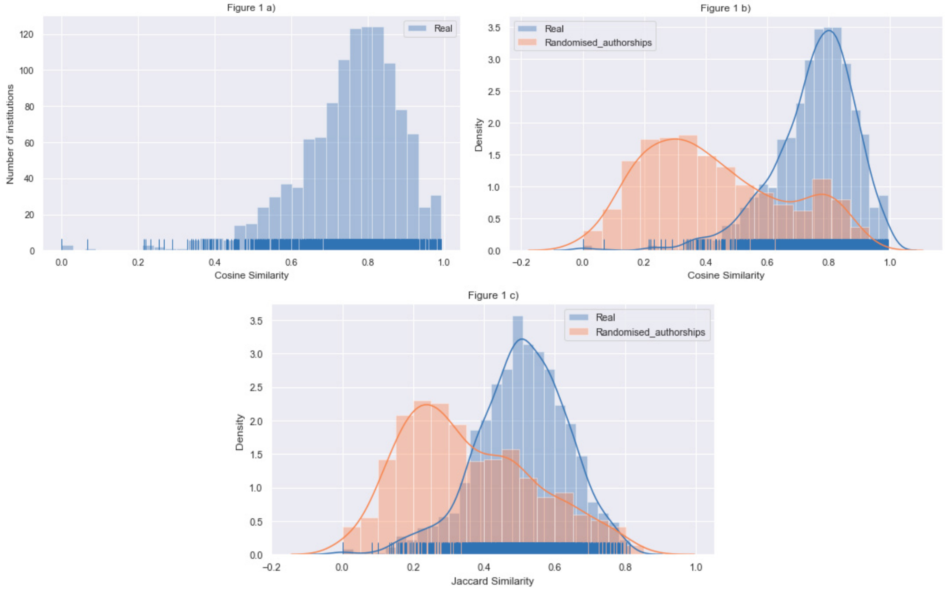


Figure 1. Cosine similarity distributions.

One potential explanation for the patterns observed in Fig. 1 is that generalist universities, already working in most micro-fields, will attract newcomers already working in common micro-fields. More specialized or smaller universities, working in more focused sets of micro-fields, may exhibit stronger differences in the comparison between newcomers and natives. This extent is explored in Fig. 2, where it is visible how indeed universities that work in very large sets of micro-fields tend to have higher cosine similarity values, with newcomers and natives being more similar, while smaller or more focused universities exhibit a larger variability (i.e., some have thematically similar newcomers, others however attract newcomers that work on more differentiated profiles at the universities).

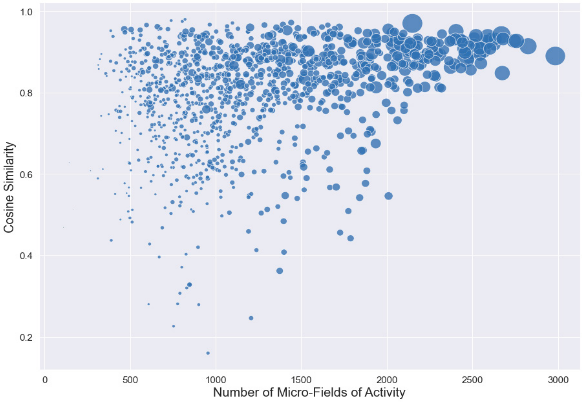


Figure 2. University cosine similarity measures of Leiden ranking universities – newcomers vs. natives.

Note: only universities with at least 100 researchers are considered in the plot; the size of the circles represents the size of the university workforce.

Discussion and conclusions

The results in this paper should be seen as a preliminary first step in order to better understand innovation patterns and the underlying recombination processes taking place at universities. There are two main conclusions that can be drawn from this analysis that can help pave the way for future research into the thematic dynamics of mobile researchers. First, universities seem to generally attract new researchers who continue to work in the same areas of the native researchers. Second, there is some relationship between the breadth of topics a university is already developing, and the ability of newcomers to substantially differentiate the profile of their new universities. In more generalist universities the potential effect of newcomers on the breadth of topics at the universities is lower, while the role of newcomers in differentiating the university profile seems more possible in specialized universities.

The two conclusions above point to some preliminary recommendations in the study of how mobile researchers could influence the thematic profile of their receiving universities: 1) the approach proposed here is more relevant to smaller and specialized universities (e.g. technical universities, medical schools, etc.), where a greater diversity in the (dis)similarities between the profiles of newcomers and natives in these universities is observed, therefore opening the possibility to analyse in which cases newcomers play a more relevant role to enlarge or differentiate the original thematic profile of their new institutions; 2) for generalist universities, it is more advisable to study them breakdown by thematically focused units (e.g. departments, faculties, institutes or research groups), instead of studying them as a whole, since only then it would be possible to assess the potential role of mobile researchers to transform, or give continuation, to the thematic profiles of the units.

It is important to highlight some of the limitations of our current study. There are two elements that will need to be further explored in future research. First, we study the output of the university once the newcomers have already joined the university, but we do not compare it with the previous thematic profiles of the newcomers. It may just be that newcomers will adapt their publishing and scientific interests to their new institutions. That is, newcomers simply ‘do in Rome what the Romans do’. Another possibility is that there may be a general tendency for universities to hire researchers with similar capabilities to their current knowledge base, or that researchers prefer to move to universities established in their own research areas. Neither of these two elements is directly measured in this paper but rather we point out the possible existence of *adaptation* and/or *homophily* mechanisms based on this preliminary evidence. We plan to address these mechanisms in the future by measuring the similarities between newcomers and natives before newcomers acquire their new affiliations. The aim is to systematically determine whether mobile researchers have already worked on the topics of their new universities, or whether they have adapted their profiles after they have moved. This could provide policy makers with advanced analytical tools to unveil the latent thematic capacity of newcomers as well as to assess the ability of newcomers to influence (or not) the thematic profile of their institutions. Such information has the potential to support and evaluate university (and its units) policies on their mobility, recruitment, and talent attraction strategies, particularly in relation to their thematic profiling.

Finally, we have not yet explored the role of collaboration in our analysis, as collaboration can be the most common approach for newcomers and natives to converge on their thematic profiles. Thus, it may very well be that newcomers influence natives to work together in the new topics they bring (Ganguli, 2015; Stephan & Levin, 2001), or vice versa; but if the level of collaboration is high between natives and newcomers, the thematic differences between would also be small. In future research we will therefore also account for the effect of collaboration in our analysis.

References

Franzoni, C., Scellato, G., & Stephan, P. (2018). Context factors and the performance of mobile individuals in research teams. *Journal of Management Studies*, 55(1), 27-59. https://doi.org/10.1111/joms.12279

Ganguli, I. (2015). Immigration and Ideas: What Did Russian Scientists "Bring" to the United States? *Journal of Labor Economics*, 33(S1), S257-S288. https://doi.org/10.1086/679741

Gertler, M.S. (2003). Tacit knowledge and the economic geography of context, or the undefinable tacitness of being (there). *Journal of economic geography*, *3*(1), 75-99. https://doi.org/10.1093/jeg/3.1.75

Macháček V., Ferreira M.R., Robinson-García N., Srholec M., Costas R. (2020). Where do scholars move? Measuring the mobility of researchers across academic institutions. September 3rd, 2020 in Leiden Madtrics Blog.

OECD. (2010). *Measuring Innovation: A New Perspective*. OECD Publishing.

Robinson-Garcia, N., Sugimoto, C.R., Murray, D., Yegros-Yegros, A., Larivière, V., & Costas, R. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics*, 13, 50-63. https://doi.org/10.1016/j.joi.2018.11.002

Sugimoto, C.R., Robinson-García, N., Murray, D. S., Yegros-Yegros, A., Costas, R., & Larivière, V. (2017). Scientists have most impact when they're free to move. *Nature*, 550(7674), 29. doi: 10.1038/550029a

Stephan, P.E., & Levin, S.G. (2001). Exceptional contributions to US science by the foreign-born and foreign educated. *Population Research and Policy Review*, 20, 59–79. https://doi.org/10.1023/A:1010682017950

Vaccario, G., Verginer, L., & Schweitzer, F. (2020). The mobility network of scientists: Analyzing temporal correlations in scientific careers. *Applied Network Science,* 5(1), 1-14. https://doi.org/10.1007/s41109-020-00279-x

Waltman, L., & Van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378-2392. https://doi.org/10.1002/asi.22748

The Software Engineering Observatory Portal

Mívian M. Ferreira[1], Bruno L. Sousa[1], Kecia A. M. Ferreira[2] and Mariza A. S. Bigonha[1]

[1] *mariza, bruno.luan.sousa, mivian.ferreira@dcc.ufmg.br*

Federal University of Minas Gerais, Dept. of Computer Science, Minas Gerais (Brazil)

[2] *kecia@cefetmg.br*

Federal Center for Technological Education of Minas Gerais, Dept. of Computing, Minas Gerais (Brazil)

Abstract

Software Engineering is a 52 years vast field of knowledge. Along the time, a massive quantity of techniques, methods, and tools have been proposed by a vast academic and industrial community, involving a large spectrum of subjects. Aiming to compile the body of knowledge of Software Engineering, we created "The Software Engineering Observatory Portal". The portal allows analyzing software engineering evolution and researchers' contributions through data visualization. So far, the portal relies on data of the International Conference on Software Engineering (ICSE) and IEEE Transactions of Software Engineering (TSE), the premier venues in the area. A video explaining the usage of the portal is available at **https://youtu.be/57Y80z9ymVw**. The portal is available at **https://mivianferreira.github.io/docs/TheSoftwareEngineeringObservatoryPortal**/.

Introduction

Over the five decades of Software Engineering, a great deal of knowledge was produced. Some initiatives have been done to compile knowledge about Software Engineering and its evolution. An essential contribution in this context is the Guide to the Software Engineering Body of Knowledge (SWEBOK) of the IEEE Computer Society (Bourque et al., 2002; IEEE Computer Society et al., 2014). In a lecture at ICSE 2018, Briand Radell, the president of the NATO Software Engineering Conference (1968-1968), highlighted the importance of knowing Software Engineering history. He made a five-decade retrospective of the area and declared: "So I hope you'll forgive me for choosing to focus primarily on historical issues, but I hope these observations encourage at least some of you to pay a little more attention to the past". Besides knowing the past, it is essential to analyze the evolution of an area of human knowledge, including Software Engineering. Such analysis may aid in answering questions such as: *What subjects the academic community has mostly considered? How much effort has been made in research on such subjects? Which papers have been most cited by researchers? Who are the researchers that have extensively contributed to Software Engineering advances? Which subjects have been mainly investigated currently?*

We constructed a web-based tool called "The Software Engineering Observatory Portal" that aims to provide resources for Software Engineering evolution analysis through data visualization. So far, we based the portal on works published in the premier Software Engineering venues: the International Conference of Software Engineering (ICSE) and the IEEE Transactions on Software Engineering journal (TSE). We retrieved data of research papers published in Transactions on Software Engineering from 1975 to 2018, accomplishing 3,357 papers. From ICSE, we gathered data of papers published from 1988, the first year IEEE Xplorer provides data of the conference, to 2018. As far as we know, this is the first initiative to compile historical data on Software Engineering with data visualization techniques. The audience of "The Software Engineering Observatory Portal" is the Software Engineering community as a whole. The portal may grow in many ways, such as: by providing other kinds of data visualizations, by considering data from other venues, and by providing updated data. Besides, the portal may be extended to other areas since the interest on investigating academic research status is global (Ioannidis et al., 2020). A video explaining the usage of the portal is available at **https://youtu.be/57Y80z9ymVw**. The portal is available at **https://mivianferreira.github.io/docs/TheSoftwareEngineeringObservatoryPortal**/.

This paper presents "The Software Engineering Observatory Portal", its main features and its construction. Moreover, this article details how the data were retrieved, describes and exemplifies the visualizations provided by the portal, and presents the conclusion of the work.

Method

We built the database applied in portal by retrieving the metadata of ICSE (data from 1988 to 2017) and TSE (1975-2017) documents. To do so, we constructed a web crawler. We collected only documents available in the IEEE Xplorer digital library, as this library presents a complete metadata structure about ICSE and TSE documents. To include a document returned by the crawler in the database, we used the following criteria: full articles, short papers, research in progress, and works available in electronic format. We carried out the exclusion of documents from the database according to the following criteria: tutorials, panels, lectures, and keynote talks; call for workshops and symposium; proceedings and round tables; presentation of sessions and tracks. We stored the collected data in a CSV file that has the following fields: title of the work; name of the authors (separated by semicolons); affiliation of the authors (separated by semicolons); year of publication; the number of downloads; keywords of the authors; keywords IEEE, NON-INSPEC or INSPEC; publication venue (ICSE or TSE).

Data Pre-processing

After retrieving the documents, we performed a pre-processing stage of the data. At this stage, we carried out the following steps:

1. Data pre-processing: we removed the documents that met the exclusion criteria and mined the keywords not returned by the web crawler. To do so, we performed a manual inspection of the articles.

2. Keywords standardization: in this step, we performed a manual inspection of the data and joining keywords with the same semantic meaning as a single keyword. This step is important because the keywords are used in the portal to classify the subjects investigated by each work. Not performing this step will introduce bias in the data. We constructed a tool to automate this step.

3. Disambiguation of authors' names: the name of an author may appear in diverse ways in the papers. This step is essential to avoid introducing bias to the authors' data. We performed a semi-automatic disambiguation process by implementing scripts to pre-process the data. We intend to develop an automatic disambiguation process to make the portal update faster and less laborious.

4. Papers results classification: the portal provides visualization based on the kind of software engineering approach presented in the works. This step consisted of setting a category for each study by reading its abstract. We used the categories defined by Mary Shaw (Shawn, 2013) to classify the studies' results. The first and the second authors carried out this classification. After that, the results were discussed among the four authors to mitigate threats to this study. This process took a significant amount of time. Due to this, so far, the portal only provides such data for ICSE papers.

Visualization Techniques

We used D3.js, CanvasJS library, Javascript, and HTML to construct the portal. It provides five types of data visualization, described as follows.

- Data Table: we chose this visualization because it is considered the most appropriate to identify the most cited ICSE and TSE Engineering papers since their titles are usually long. Besides, we used the table to present other important data related to the articles, such as authors, number of citations, year of publication, and link to the article in the *IEEE Xplorer* database. The table is presented as a search engine, ordering (increasing and descending) of all fields, paging, and choice of entries viewed per page, ensuring a good usability.

- Dispersion Plot: we used the dispersion plot to show data of authors whose contributions have a significant impact. We consider that the author's contribution is determined by the number of citations of an author and the number of articles published by that author – thus establishing the need to use the dispersion plot. To determine the authors whose contributions have a high impact, we represented each author by a circle, whose area is proportional to the number of citations, i.e., the greater the number of citations of an author, the greater the circle's size. The x and y axes represent the number of citations of an author and the number of articles published by an author. To facilitate the visualization of the data, we subdivided the authors into three categories: All, with all authors; Top-10, with only the ten authors with the highest number of citations; and Top-100, with the 100 authors with the highest number of citations. Besides, there is a tool-tip in each of the circles where we displayed the author's name, number of citations, and articles.

- Keywords Cloud: this visualization shows a cloud with the subjects investigated in the area. In this cloud, the larger the font size of a keyword label, the larger the number of works considering the related subject. This visualization allows researching keywords according to a time frame chosen by the user. For instance, the user may want to see the cloud of works published from 2015 to 2018. When the user move the cursor on the word, the portal shows the corresponding number of papers. Hence, when the user clicks the word, the portal shows a data table with all the works in which the chosen keyword appears.

- Overlapping areas: we used this visualization to show the amount of research done on the top subjects that have been investigated in Software Engineering. In this visualization, each layer represents a keyword. The x and y axes represent the years used for data analysis and the number of occurrences of the keywords, respectively.

- Pareto Chart: we developed this visualization with CanvasJS. The portal provides a Pareto's (80/20) chart of the following data: number of citations/article, number of citations/author, and number of articles/author.

Results

In this section, we exemplify the use of "The Software Engineering Observatory Portal".

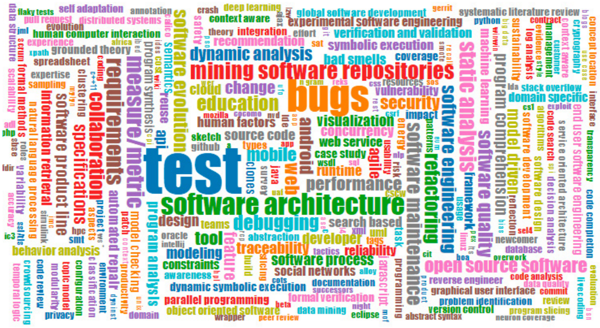
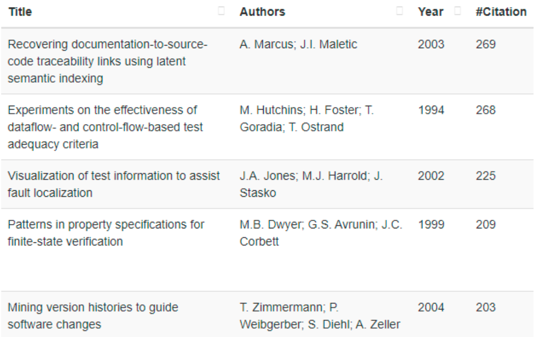


Figure 1. Keywords cloud of ICSE with data from 2015 to 2018.

The most investigated subjects in a given period.
The user selects the range of years and the portal shows the most investigated subjects in a keyword cloud. Figure 1 shows the keywords cloud of 1988 until 1990.

The most cited papers
The number of citations of a paper is an indicator of its relevance. The portal provides a table with data of the publications, including the number of citations. Figure 2 shows an example of this table.



Title	Authors	Year	#Citation
Recovering documentation-to-source-code traceability links using latent semantic indexing	A. Marcus; J.I. Maletic	2003	269
Experiments on the effectiveness of dataflow- and control-flow-based test adequacy criteria	M. Hutchins; H. Foster; T. Goradia; T. Ostrand	1994	268
Visualization of test information to assist fault localization	J.A. Jones; M.J. Harrold; J. Stasko	2002	225
Patterns in property specifications for finite-state verification	M.B. Dwyer; G.S. Avrunin; J.C. Corbett	1999	209
Mining version histories to guide software changes	T. Zimmermann; P. Weibgerber; S. Diehl; A. Zeller	2004	203

Figure 2. Table with the data of the most cited ICSE papers.

The researchers that have mainly contributed to Software Engineering advance. The portal provides visualizations of data about the number of articles and the number of citations by author. The visualization is a Dispersion plot, as shown in Figure 3. Figure 4 shows the Pareto chart of citations per article of ICSE. When the user moves the mouse over the curve, the portal shows the cumulative number of articles corresponding to the percentage of citations.

The most investigated subjects along with Software Engineering evolution. The 11 most researched topics in ICSE were: test (236 occurrences); bugs (147 occurrences); software architecture (132 occurrences); measure/metric (87 occurrences); requirements (85 occurrences); education (74 occurrences); empirical study (59 occurrences); software quality (56 occurrences); software process (54 occurrences); software engineering (53 occurrences); and reuse (52 occurrences). We used an overlapping areas chart to show the evolution of the number of studies on those subjects. Figure 5 shows the chart for ICSE.

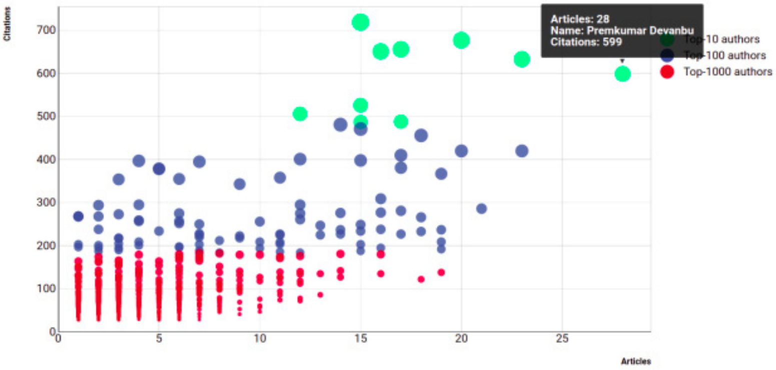


Figure 3. Dispersion plot with data of the 1000 most cited ICSE papers.

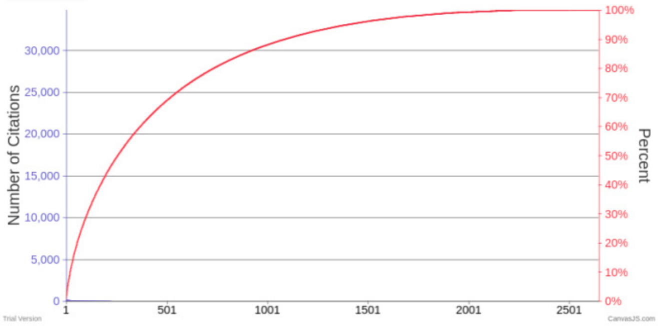


Figure 4. Pareto chart with number of citations of ICSE papers.

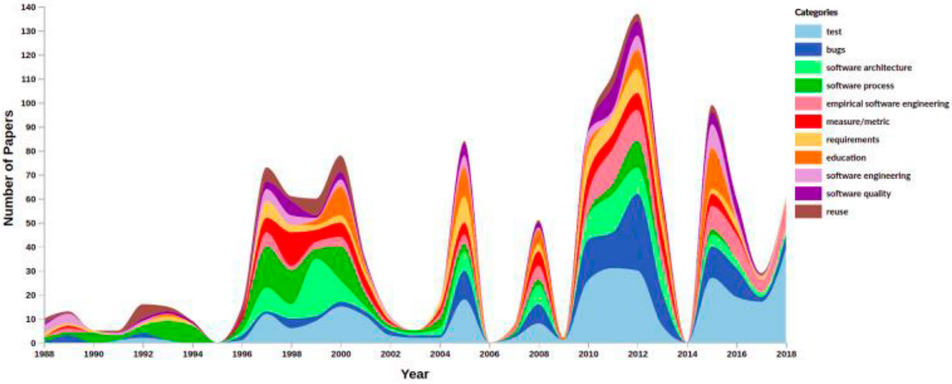


Figure 5. Overlapping area showing the evolution of number of papers considering the top-10 most investigated topics - data of ICSE.

Related Work

Visualization is a powerful resource to aid bibliometrics analysis. The study of Muñoz et al. (2020) shows that there are many tools for this purpose. The main difference between those tools and our portal is that the portal is an on-line and interactive tool containing pre-processed data of an specific area, Software Engineering. Besides, the visualizations resources provided by the portal are also different. Liao et al. (2018) used data visualization charts, such as co-authorship networks and keywords density map, to analyze bibliometrics data in Medicine. The visualization they used are different from the ones we applied. Moreover, they did not construct a portal as we did. Other initiatives have been taken to compile the body of knowledge on Software Engineering, such as the SWEBOK (Bourque et al., 2002; IEEE Computer Society et al., 2014) and Boehm's work (Boehm, 2006). However, as far as we know, this work is the first one that compiles Software Engineering historical data by applying visualization techniques. We used the Overlapping Area chart in our work on Software Engineering evolution considering ICSE data (Sousa et al., 2019).

Conclusions

This paper presents a web-based tool called "The Software Engineering Observatory Portal" that provides resources for analyzing the evolution of Software Engineering, as well as data of contributions, number of citations, and subjects primarily investigated in the area. The portal applies data visualization and relies on papers publishes in ICSE and TSE, the premier software engineering discussion forums. The analysis provided by the portal is of interest to the Software Engineering community as a whole, especially the academia. As part of the pre-processing data was done manually and it is exceptionally laborious, we are developing an automatic tool to construct a thesaurus of Software Engineering and adapting the gathering data process to the new structure of IEEE Xplorer. Besides, we are investigating a proper approach to disambiguate authors' names. We believe that the portal could be applied to other Computer Science disciplines whose publications are available in IEEE Xplorer.

Acknowledgments

This work was supported by CAPES, CNPq and CEFET-MG.

References

Boehm, B. (2006). A view of 20th and 21st century Software Engineering. In *Proceedings of the 28th International Conference on Software Engineering (ICSE '06)*, p. 12–29.

Bourque, P., Lavoie, J.-M., Lee, A., Trudel, S., Lethbridge, T.C, *et al.* (2002). Guide to the software engineering body of knowledge (SWEBOK) and the software engineering education knowledge (SEEK)-a preliminary mapping, In: *10th International Workshop on Software Technology and Engineering Practice*. IEEE Computer Society, pp. 8–8.

IEEE Computer Society, Pierre Bourque, and Richard E. Fairley. (2014). Guide to the Software Engineering Body of Knowledge (SWEBOK(R)): Version 3.0. *IEEE Computer Society Press.*

Ioannidis, J.P.A., Boyack, K.W. & Baas, J. (2020). Updated science-wide author databases of standardized citation indicators. *PLoS Biol*ogy, 18(10): e3000918.

Liao, H., Tang, M., Luo,L., Li ,C., Chiclana, F. & Zeng, X. 2018). A bibliometric analysis and visualization of medical big data research. *Sustainability*, 10(1), 166. https://doi.org/10.3390/su10010166

Muñoz, J.M., Viedma, E.H., Espejo, A.S. & Cobo, M.J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. El Profesional de la Información, 29(1), https://doi.org/10.3145/epi.2020.ene.03.

Shaw, M. (2003). Writing good software engineering research papers: Minitutorial. In:*25th International Conference on Software Engineering (ICSE)*, 726–736.

Sousa, B.L.,Ferreira, M.M., Ferreira, K.A.M. & Bigonha, M.A.S. (2019). Software engineering evolution: The history told by ICSE", in XXXIII *Brazilian Symposium on Software Engineering*, 17-21.

Tracing epistemic effects of research governance: a mixed-method approach

Thomas Franssen[1], Thed van Leeuwen[2], Ingeborg Meijer[3] and Ismael Rafols[4]

[1] *t.p.franssen@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies, Kolffpad 1, 2333 BN Leiden (The Netherlands)

[2] *leeuwen@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies, Kolffpad 1, 2333 BN Leiden (The Netherlands)

[3] *i.meijer@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies, Kolffpad 1, 2333 BN Leiden (The Netherlands)

[4]*i.rafols@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies, Kolffpad 1, 2333 BN Leiden (The Netherlands)

Abstract

One of the challenges in the sociology of science is to increase our ability to describe and analyze the epistemic effects of research governance. One reason for our current limitations in doing so is an epistemic divide between different fields that study (parts of) this question. Science policy studies focus on developments in research governance but lack interest in its epistemic effects. Studies of epistemic properties of research often lack interest in the institutional environment in which research is conducted. Moreover, scientometric methods are not often combined with qualitative methodologies. This lack of mixing methods hampers progress because for the study of the epistemic effects of research governance it is often necessary to employ a range of methods at the same time. This research-in-progress paper reports on an in-depth study of a biomedical research school in the Netherlands. Over the last 20 years, spurred on by science policy and external evaluations, the school has attempted to bring about more integration in its research program. We aim to offer a mixed-method approach that can be used to advance the sociology of science its ability to trace epistemic effects of attempts to steer research content in public research organizations.

Introduction

One of the challenges in the sociology of science is in the limited ability we currently have to describe and analyze the epistemic effects of research governance. One reason for this is the split between fields that study parts of this question. Science policy studies analyze the dynamics of research governance but do not study its epistemic effects. Fields in which the epistemic properties of research are central, notably science and technology studies, often lack interest in the institutional environment in which research is conducted (Gläser & Laudel, 2016).

The methodological divide that emerged in the 1980s (Wyatt et al., 2017) between those that employ quantitative methodologies and those that employ qualitative methodologies makes solving this all the more difficult. This methodological divide hampers progress because for the study of the epistemic effects of research governance it is necessary to employ a range of methods at the same time. In recent years, there is a renewed interest in the integration of perspectives and methods (e.g. see Wallace and Rafols, 2018) and we aim to contribute to this research agenda.

Our substantive interest in this paper is the epistemic effects of research governance within public research organizations. Public research organizations have increased their strategic capacity (Thoenig & Paradeise, 2016) in response to changes in science systems that introduced new external pressures and demands. One way in which research governance takes place within public research organizations is through the development and implementation of strategic goals. The advantage of focusing on the epistemic effects of strategic goals is that these are well documented in, for instance, self-evaluation documents and mission statements. They can also be discussed with staff and management in interviews. We regard the analysis of the implementation of strategic goals in public research organizations as interesting test cases to establish new connections between sociological concepts used to describe epistemic effects and bibliometric measures used to operationalize these concepts (Gläser & Laudel, 2001).

In this research-in-progress paper we present an ongoing in-depth study of research governance in a single biomedical research school in the Netherlands. We draw on documents, in particular those written in the context of three external evaluations conducted in the time span of 20 years, have gathered publication data which was validated by the research school itself, and are doing interviews with staff members and the management team of the research school. Drawing on this variety of sources allows us to develop links, in a systematic way, between epistemic changes in the research conducted at the research school and the steering mechanisms employed by the management team of the research school to implement its strategic goals.

Our approach involves three steps. In the first step we analyze the strategic goals at the level of the research school over time. The second step consists of an analysis of changes in the research content of research conducted employing bibliometric methods. Here we seek to test the hypotheses formulated in step one and employ a range of bibliometric methods understood as 'partial indicators' (Martin & Irvine, 1983) in search of the expected effect. The third step, is to unpack the mechanisms that produced the epistemic effect found bibliometrically. This step consists of identifying the steering mechanisms employed in the research school from interviews and document analysis and, where possible, relate them to the epistemic changes observed with bibliometrics.

Case description: Strategic goals of a biomedical research school

The biomedical research school we study was established in the 1980s and soon rose to a prominent position, nationally and internationally, in its subfield. The school is comprised of three research units that themselves consist of principal investigators (PIs) groups. Each unit contains somewhere between 5 and 10 PI-groups, but the number of groups changes over time.

An important external pressure that drives organizational change in the research school is the Dutch research evaluation cycle. Every 6 years a research evaluation is conducted by an external review committee which is followed by a mid-term evaluation 3 years on and preceded by a self-evaluation the outcome of which is send to the external review committee as a preparation. Conducting the evaluation is mandatory for all research schools and departments in Dutch public research organizations. The evaluation report is usually basis for discussion between a management team of a school or department and senior level management at the university (for instance the dean of the faculty).

Conducting the self-evaluation increases the need for strategic capacity within the research school. Similarly, the external evaluation and its cyclical nature increases the need to reflect on and act on the outcomes of the evaluation and report on these actions and their results in the subsequent evaluation. It thus pushes a management team to continuously change (or improve) the organization by, for instance, establishing new strategic goals and a strategic agenda, new organizational processes and so on. As an example, consider the following quote from the introduction to the self-evaluation report of the period 1999-2006:

> "The advise of the ERC [External Review Committee, TF] is seen as a major tool for the Board and Management of [research school] to sustain and improve the quality of their research and education programs and to give foundation to strategic decision-making procedures. (…) [W]e have chosen to describe our strategic goals for the next 5 years and hope that the External Research Committee will give us feedback that will help us to further improve and clarify our strategy." (self-evaluation report 1999-2006)

The biomedical research school we study went through three evaluations in the last 20 years (evaluations of the periods 1999-2006, 2007-2012, 2013-2018). A constant factor in these evaluation reports is reflection on the coherence of the research profile of the school. The research in the school covers all main topics concerning the biomedical subfield it is part of and argues this diversity makes it unique. Yet it also aims to achieve coherence. This is discussed both at the level of each research unit (that individually have to be coherent) as well as across the research units. In the first evaluation report we can read that one of the strategic goals is 'To maintain a strong and coherent research program'. Later on in the document this goal is expanded upon:

> A recent change is that all three [research units] have defined their research programs around larger (sub) themes, further strengthening the coherence of the research and increasing the critical mass, in line with the research strategy of the Faculty and the research strategy 'Focus and Mass' of the KNAW [Royal Dutch Academy of Sciences, TF]. (self-evaluation report 1999-2006)

As the quote shows, the school's strategic goal for a coherent research program is implemented through the development of larger subthemes, composed of PI-groups, within the three research units. The strategy is legitimated by connecting it to the notion of critical mass which is one of the pillars of Dutch science policy since the 1980s (e.g. See Franssen, 2020). This organizational development was continued over the next 13 years and in the last evaluation, concerning the period 2013-2018 the structure of the research school is discussed again. The self-evaluation describes the research units as well governed and interconnected (both internally and between them) entities:

> All three [units] involve basic as well as clinical programmes and are led according to a shared governance principles (…) executed by the [unit] leader together with one or more basic and clinical scientists from the [Unit]. This shared governance system enables shared responsibility for the scientific progress of programmes, for linking activities and seeking collaborations between PIs and [units] (…) Due to the increasingly complex and multidisciplinary [biomedical subfield] research objectives, [research school's] new structure facilitates much needed inter[unital] and translational research. (self-evaluation report 2013-2018)

As this quote shows the biomedical problems the school is focused on require different disciplinary knowledges to be connected. The need for connectivity is stressed by pointing to collaboration across units and between basic and clinical research. In the evaluation report this self-evaluation is however criticized for not describing how coordination and collaboration are organized between the research units. The evaluation argues for this to be developed further:

> In the self-evaluation report there is only a sparse description of how – and by which means – the coordination and collaboration between [units] – [unit 1], [unit 2] and [unit 3] - takes place. During the site-visit it was described that the Executive Boards meet on a monthly basis, however, a more detailed description of how collaboration between the three [units] are facilitated by the Executive Board is needed, both strategically and on a daily basis. (evaluation report 2013-2018)

What is alluded to in these evaluation reports is, we argue, a concern with increasing knowledge integration both within each of the three research units as well as across these research units. The concept of integration has a long history in science studies and has primarily been studied in relation to integration of scientific communities (Gläser & Laudel 2001). In our case knowledge integration is a policy goal (see Luukkonen & Nedeva, 2010) within a public research organization in order to develop a more coherent research profile. As a first step, in this work-in-progress proceeding, we propose two hypotheses and their operationalization, to analyze the implementation of this strategic goal:

Hypothesis one: diversity within research units goes down due to the increase in integration of the research program of each research unit.

> This hypothesis is tested through measuring the cognitive diversity within research units, operationalised as a diversity measure of the distribution of publications in each research unit across micro clusters.

Hypothesis two: diversity between research units goes down due to increased integration between research units.

> This hypothesis is tested through measuring cognitive similarity between units, as shown by topics of publications and referencing practices. This is operationalised in two measures:
>
> - Cosine similiarity of the microclusters of publications between research units.
> - Cosine similiarity of references of publications between research units.

Using these indicators, we find evidence of the dynamics of integration between research units. This data will be combined with insights from interviews and documents to explain the mechanism behind increasing knowledge integration within the public research organization.

Data and Methods

We have composed a dataset of all publications of the staff of the research school between 2000 and 2019 which was validated by the administrative personal of the university medical center to which the research school belongs. We have subsequently divided the publications into three periods in line with the evaluation cycles that the school underwent, 2001-2006, 2007-2012, 2013-2019.

These publications were linked to CWTS in-house version of Web of Science. Each publication in the Web of Science is assigned to a cluster using the CWTS article-level algorithm (Waltman & van Eck, 2012). At the lowest level of granularity it consists of some 4000 microclusters. For each publication from the research school the microcluster it belongs to was determined, much like one would use the Web of Science journal categories. An upside to using the microclusters is that these are based on citation links between papers rather than journal-based. Moreover, we circumvent the category 'multidisicplinary journals' which is hard to interpret and holds many publications of the research school especially in the third period.

To assess cognitive diversity within a unit, we use conventional indicators of diversity (variety, Shannon entropy, and Shannon evenness), as explained in detail in Rafols et al. (2012). The cosine similarity is used to assess the degree of similarity between vectors representing different units, where a vector per unit describes, the number of publications of a unit across clusters. This indicator is repeated using a vector where each dimension is a reference with the number of times a reference is cited by paper. The relative similarity due to collaborations is computed as the similarity between units including co-authored papers between units minus the similarity without co-authorships, divided by the similarity between units with co-authorships.

Results

Hypothesis 1: Cognitive diversity within units

We use number of microclusters (variety) and Shannon as measures of the diversity of topics to which publications belong. Given that the number of publications is increasing, we also compute Shannon diversity to measure the balance of topics. Against expectations of integration, we find that the measures of diversity increase, whereas the measure of balance is stable since the second period. These results shows that topical diversity within research units is not reduced as a consequence of a strategic decision to increase coherence. In the next step of this paper we will use other measures of integration, such as bibliographic coupling between PI-groups within research units, to assess the extent to which the implementation of strategic goals has led to an increase in integration without affecting diversity as measured through microclusters here.

Table 1. Diversity measures: Number of microclusters, Shannon entropy and Shannon evenness.

Variety (number of microclusters)	**2001-2006**	**2007-2012**	**2013-2019**
Research unit 1	185	232	288
Research unit 2	81	133	186
Research unit 3	311	404	412
Shannon entropy			
Research unit 1	4.21	4.53	4.50
Research unit 2	3.24	3.86	4.11
Research unit 3	4.87	5.29	5.23
Shannon evenness			
Research unit 1	0.81	0.83	0.79
Research unit 2	0.74	0.79	0.79
Research unit 3	0.85	0.88	0.87

Hypothesis 2: Cognitive similarity between units
In order to estimate the similarity in the research content across the three research units, we compare the distribution of their publications across the microclusters. The analysis shows (table 2) that there is an increasing degree of similarity between the three units in terms of microclusters. This result is confirmed using references as units for comparison (table 3). This suggests that the strategic goal to increase integration between research units has been successful. But what mechanism might have been used to foster integration?

Table 2. Cognitive similarity according to micro cluster distribution

All publications in units	**2001-2006**	**2007-2012**	**2013-2019**
Research units 1-2	0.05	0.07	0.13
Research units 1-3	0.12	0.21	0.25
Research units 2-3	0.07	0.21	0.24
Without collaborations between units			
Research units 1-2	0.04	0.03	0.08
Research units 1-3	0.11	0.09	0.08
Research units 2-3	0.07	0.09	0.10
Relative similarity due to collaborations			
Research units 1-2	10%	61%	36%
Research units 1-3	9%	55%	70%
Research units 2-3	5%	57%	56%

One possible explanation is increased collaboration, in terms of co-authorship, between research units. Co-authorships between units increase from 2% of the total in the first period to 18% in the second period to 23% in the third period. Publications shared between units increase similarity of the research profile of the three units. In table 2 and 3 we have included the same measures of cognitive similarity taking out all papers that were collaboration betweens research units (and thus present in as publications in both sets of publications). We also measure the ratio to which cognitive similarity can be ascribed to these collaborations. This shows that in the second and third period a large part of cognitive similarity, especially when measured through reference similarity, can be assigned to collaborations. Looking at microclusters a large part of cognitive similarity is also due to collaboration but overall 40% or more percent cannot be attributed to collaboration. Using these two indicators for cognitive similarity shows that while the point in the same direction, they measure cognitive similarity in a slightly different way.

Table 3. Cognitive similarity according to references

All publications in units	**2001-2006**	**2007-2012**	**2013-2019**
Research units 1-2	0.028	0.067	0.121
Research units 1-3	0.062	0.157	0.218
Research units 2-3	0.041	0.205	0.226
Without collaborations between units			
Research units 1-2	0.021	0.009	0.037
Research units 1-3	0.050	0.041	0.050
Research units 2-3	0.037	0.047	0.053
Relative similarity due to collaborations			
Research units 1-2	25%	87%	70%
Research units 1-3	18%	74%	77%
Research units 2-3	9%	77%	77%

Conclusion

This proceeding aims to illustrate how bibliometrics can show the epistemic changes produced by strategic decisions in research management with an example from a biomedical school. These initial results point to collaboration in publications as an important mechanism for increased knowledge integration in the research school. We will therefore extent our analysis of collaboration, using both qualitative and quantitative data and methods to understand who is collaborating with whom and why these researchers are more prone to collaborate in period 2 and 3.

References

Franssen, T. (2020). Research infrastructure funding as a tool for science governance in the humanities: a country case study of the Netherlands. In K. Cramer & O. Hallonsten (Eds.), *Big science and research infrastructures in Europe*. Cheltenham: Edward Elgar.

Gläser, J. & Laudel, G. (2001). Integrating scientometric indicators into sociological studies: methodical and methodological problems. *Scientometrics*, 52, 411-434.

Gläser, J. & Laudel, G. (2016). Governing science: How science policy shapes research content. *European Journal of sociology*, 57, 117-168.

Luukkonen, T. & Nedeva, M. (2010). Towards understanding integration in research and research policy. *Research Policy*, 39, 674-686.

Martin, B, & Irvine, J. (1983). Assessing Basic Research. Some partial indicators of scientific progress in radio astronomy. *Research Policy,* 12, 61–90.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P. & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research policy*, 41, 1262-1282.

Thoenig, J. & Paradeise, C. (2016) Strategic Capacity and Organisational Capabilities: A Challenge for Universities. *Minerva,* 54, 293–324.

Wallace, M. L. & Ràfols, I. (2018). Institutional shaping of research priorities: A case study on avian influenza. *Research Policy*, 47, 1975-1989.

Waltman L. & van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system of science, *Journal of the American Society for Information Science & Technology*, 63, 2378-2392

Wyatt, S., Milojević, S., Park, H. & Leydesdorff, L. (2017). The intellectual and practical contributions of scientometrics to STS. In U. Felt, R. Fouché, C. Miller, and L. Smith-Doerr (Eds.), *Handbook of Science and Technology Studies* (pp. 87–112), Boston, MA: MIT Press.

Using Citation Bias to Guide Better Sampling of Scientific Literature

Yuanxi Fu, Jasmine Yuan and Jodi Schneider

{fu5, jyuan25, jodi}@illinois.edu
University of Illinois at Urbana-Champaign, School of Information Sciences, 501 E Daniel St, Champaign, IL 61820 (USA)

Abstract

It is rarely possible to cite every relevant work on a topic. When controversy exists in a field, work holding the same opinion as the citing paper (i.e., homophily) is more likely to be cited. Thus, readers may inadvertently select a non-representative sample of articles to read. Here, we begin to develop a method that guides better sampling of scientific literature by designing and testing two new network metrics. The first metric, the ratio between real and expected citation counts, guides users to papers that were cited many fewer times than expected and may represent marginalized findings. The second metric, the relative evidence coupling strength, guides users to papers that may present a unique view of the field. We test our metrics on a known case of citation bias: a network of 73 papers about whether stress is a risk factor for depression. Our metrics select a cross-section of 21 papers. The intersection of the two metrics selects 3 papers that represent all 3 positions of this claim network. In future work we will test our metrics on more datasets, and we will partner with domain experts to verify whether our metrics do identify a diverse sample of research articles.

Introduction

Authors must make choices when citing related work: it is rarely possible to cite every relevant article on a topic, and sometimes even identifying all relevant work is difficult. However, citation bias arises when "data critical of or refuting claims are systematically uncited in favour of data supporting a claim" (Greenberg, 2009). Studies have found that cited articles are more likely to report statistically significant results (Urlings, Duyx, Swaen, Bouter, & Zeegers, 2019b), outlier studies (Schrag, Mueller, Oyoyo, Smith, & Kirsch, 2011), or conclusions supportive of a hypothesis (Duyx, Urlings, Swaen, Bouter, & Zeegers, 2017; Urlings, Duyx, Swaen, Bouter, & Zeegers, 2019a). When controversy exists in a field, preference was given to articles holding the same opinion as the citing paper (i.e., homophily) (Trinquart, Johns, & Galea, 2016). Citation bias benefits authors by bolstering grants and papers, making them more easily accepted. However, it creates a "filter bubble" (Pariser, 2011) for people who want to use scientific literature as impartial evidence. For example, busy decision-makers only have time to read a few articles. They may end up inadvertently choosing articles belonging to a group of papers that reach the same findings. This selection of the literature gives the impression that there are no controversies on the given topic and that the evidence is well supported by many other papers.

This paper addresses an important challenge facing many literature users: how to sample diverse research articles. In this paper we propose a new sampling method, which uses insights from citation bias studies to design new network structure metrics to guide better sampling of the literature. First, we need to identify papers that get far fewer citations than chance permits, in case they are marginalized due to contradicting the dominant view of a field. Second, we need to identify papers that do not belong to any homophily group(s), because they may hold unique views towards the scientific question. Furthermore, the method should be accessible and intuitive, so that it can easily be adopted by a wide range of users. This paper reports our work in progress on developing such a method. First, we explain our current method, along with the design of two network structure metrics. Then we apply our method to a citation network. And finally, we discuss our plans for future work.

Data

To answer a specific scientific question, decision makers gather a body of literature relevant to some topic. This set of papers (nodes) and the citation relationships among them (directed edges) form a "claim-specific network," following Greenberg (2011). A claim-specific network is a subgraph of the entire citation network, since we omit all papers that do not concern a given topic (Greenberg, 2011). Thus, the papers (nodes) in a claim-specific network all address a single research question, for example, whether reducing salt intake brings health benefit at the population level, or whether stress is a risk factor for depression. Such a network captures the communication of ideas and the establishment of belief regarding a specific scientific question (Greenberg, 2011).

The dataset used here comes from a paper that analyzed citation bias in favor of positive findings (de Vries, Roest, Franzen, Munafò, & Bastiaansen, 2016). That study constructed a claim-specific citation network between 73 primary studies, all concerning risks of developing depression after stressful life events, associated with a gene known as 5-HTTLPR. Based on the outcome each primary study reported, 24 studies were coded as "Positive", 38 studies as "Negative", and 11 studies as "Unclear." The "Positive" articles received more citations than the "Negative" articles (de Vries et al., 2016).

Metrics

In this pilot study, we will use two metrics. The first metric is the **ratio between the real and expected citation count**, which measures whether a paper is sufficiently cited. Currently, we define the expected citation count as the number of citations a paper would receive if all papers in the network published *two years or more later* cited it. This ratio is designed to select papers that would be less prominent in a purely citation-based heuristic.

To compute the ratio between the real and expected citation count for the 5-HTTLPR claim-specific citation network, we constructed two networks. First, we reconstructed the real claim-specific citation network from the published dataset (de Vries & Munafò, 2016), which has 73 nodes and 488 edges. Second, we constructed a new, simulated network, in which the nodes are identical to those in the real network, while the edges are determined by the gap between the publication years of the two articles. To be more specific, if two papers are two or more years apart, we add a directional edge from the younger paper to the older paper. The resulting simulated network has 73 nodes and 1799 edges. Currently, we compute the expected citation count of a node as its in-degree in this simulated network.

The second metric resembles the relative bibliographic coupling strength (Sen & Gan, 1983; Shen, Zhu, Rousseau, Su, & Wang, 2019). The difference is that we limit to the items within the claim-specific network, rather than the whole bibliography: thus, in the name we replace "bibliographic" with "evidence". Specifically, we define the **"relative evidence coupling strength"** as

$$\text{RECS(X,Y)} = \frac{|E_X \cap E_Y|}{|E_X \cup E_Y|}$$

where E_X and E_Y represent the set of references for X and for Y in the claim-specific network. In other words, this metric characterizes the percentage of overlapping items (i.e., $E_X \cap E_Y$) in the collective evidence set (i.e., $E_X \cup E_Y$) between the pair of nodes. This metric is inspired by an observation we made in a study of disagreement between systematic reviews, in which we found that systematic reviews that synthesized different evidence sets reached different conclusions (Hsiao, Fu, & Schneider, 2020).

Although our two metrics were inspired by studies of citation bias, their ultimate utility is to help literature users sample a diversity of research articles. Users do not need to know whether a claim-specific network is biased or to what extent to make use of them.

Our script to construct the networks and compute the network metrics can be found in GitHub (https://github.com/infoqualitylab/citation_bias_study/tree/master/ISSI_2021).

Results

Figure 1(a) shows the distribution of our first metric, the ratio between the real and expected citation counts. Citations are quite unequal among those 73 papers. One paper, Caspi 2003, received 97% of its expected citation counts, whereas 16 papers (22%) (colored in pink in Figure 2) received less than 10% of their expected citation counts. Since they were cited many fewer times than expected, those 16 papers need more attention from readers, in case they represent a marginalized view of the topic. Also, 18 papers were published in the last two years covered in the datasets (2011 and 2012), and under our current construction, their expected citation counts are zero, and therefore do not have a value for this metric. Those papers are colored white in Figure 2.

We computed the relative evidence coupling strength for each of the 2628 pairs. Figure 1(b) shows the distribution of this metric. Forty-nine pairs (1.9%) shared more than 90% of their collective evidence set and 818 pairs (31%) shared less than 10% of their collective evidence set. Since a single paper can be involved in 72 pairs, it is quite common for a paper to find another paper in the dataset with which its relative evidence coupling strength is small. What we need to pay attention to are papers that *always* have a small relative evidence coupling strength with other papers. Those are what we call "unique" papers, as they may represent unique views of the field (i.e., looking at a different set of evidence). Currently, we set an arbitrary threshold of 0.3 and found 8 papers that never participated in any pair with relative evidence coupling strength more than 0.3. Those papers are colored in pink in Figure 3 and they should be sampled too.

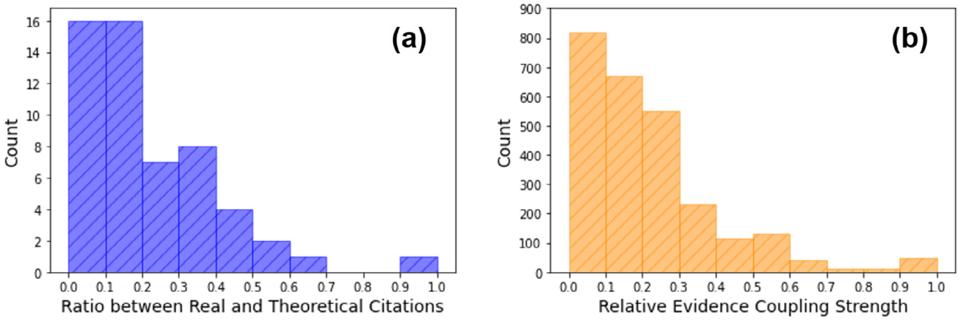


Figure 1. The distribution of the ratio between (a) the real and expected citation counts and (b) the relative evidence coupling strength for the 5-HTTLPR network

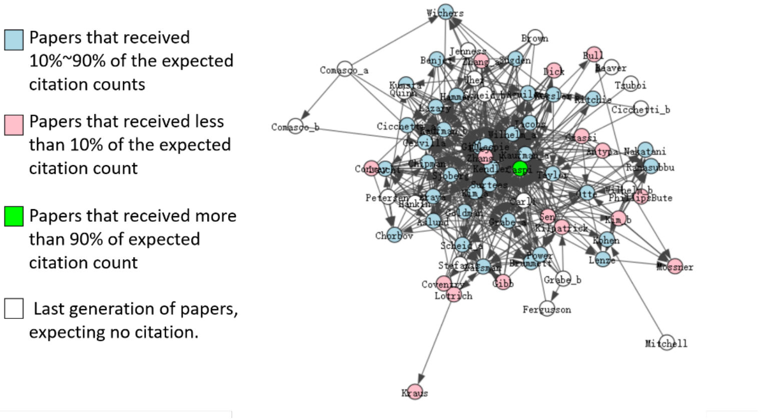


Figure 2. The ratio between real and expected citation count applied to the HTTLPR-5 network

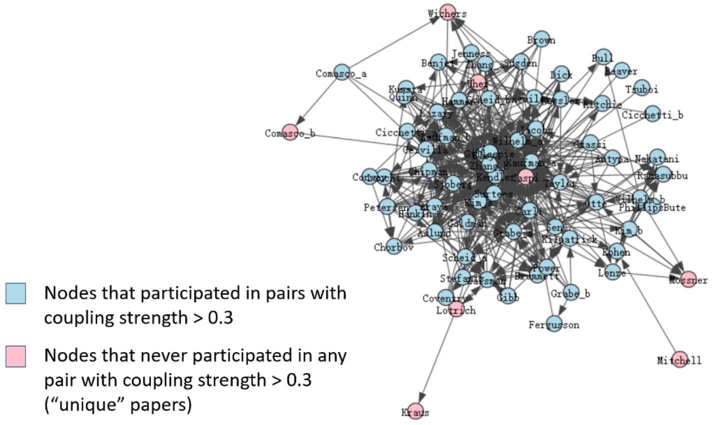


Figure 3. Relative evidence coupling strength applied to the HTTLPR-5 network

Our two metrics select different articles. As shown in Table 1, there are just 3 overlaps. Interestingly, each group selects papers with all 3 available outcomes (Positive, Negative, Unclear) – and the group selected by the intersection of the two metrics does this in just 3 papers, the minimum possible. Notably, there is only a 16.1% chance for a randomly selected set of three papers to cover all three outcomes. This finding suggests that these metrics are promising for further exploration, because the intersection included three papers with different outcomes, meeting our intention to select a diverse set of papers for readers to read.

Table 1. Papers selected by the two network metrics

Insufficiently cited papers	Dick 2007 Unclear Kilpatrick 2007 Unclear Bull 2008 Positive Zhang_a 2008 Negative PhillipsBute 2008 Negative Zhang_b 2009 Negative Kim_b 2009 Positive	Gibb 2009 Unclear Coventry 2009 Negative Grassi 2010 Negative Sen 2010 Positive Antypa 2010 Negative Conway 2010 Negative
Both	Mossner 2001 Positive Lotrich 2008 Unclear	Kraus 2007 Negative
"Unique" papers	Caspi 2003 Positive Comasco_b 2011 Unclear Wichers 2007 Negative	Uher 2011 Negative Mitchell 2011 Negative

Conclusions and Future Works

In conclusion, we proposed a method that may help users who need to select diverse papers to read to assist in their decision making. We designed two metrics inspired by previous studies of citation bias: First, the ratio between the real and expected citations identifies papers that received far fewer citations than expected, which may represent marginalized views. Second, the relative evidence coupling strength identifies papers with unique views towards a topic. We applied these metrics to a previously citation network of 73 research articles studying whether a gene increases the risk of depression after stressful events. Our sampling method resulted in partitioning the network into two groups: 21 papers that readers need to pay particular attention to, and 52 articles from which readers can choose a few to read.

Our current method has known limitations and needs evaluation and improvement in the future. First, the current expected citation count is in effect a maximum citation count (assuming a delay between publication and citation). Some of the later published papers were expected to cite as many as 55 papers, which is unrealistic and inflated the expected citation counts for other papers. Second, the relative evidence coupling strength also has low values among paper pairs whose publication years are far apart, and the method we are using now (i.e., selecting nodes that never participated in any pair with relative evidence coupling strength above a threshold value) over-selects early papers, which have a diminished chance to share reference items with papers published later.

For future work, our priority is to find domain experts to evaluate the two sets of papers and see whether the selected papers fit our expectations. Second, we need to evolve the metrics. We plan to create better algorithms to for computing more realistic expected citation counts. And we will also filter out papers selected purely due to their age. The visualization should be improved too, making it more interactive and informative. We envision an interface that allows users to zoom into one part of the network and click on a node to find information about a paper (e.g., title, DOI). Finally, the utility of our methods will need to be verified with larger-scale studies. We plan to use both published claim-specific networks (e.g., Duyx et al., 2019; Trinquart et al., 2016; Urlings et al., 2019a; Urlings et al., 2019b) as well as our own datasets constructed in collaboration with domain experts. We are currenting working with a kinesiology and community health expert and have constructed a claim-specific citation network with 439 articles relating to the effectiveness of exercise therapy for treating depression. This is larger than any claim-specific citation network that we are aware of. Those datasets will provide a larger testbed for verifying whether our method is effective in helping readers obtain a diverse sample of research articles.

Acknowledgments

This research is supported by the Campus Research Board of the University of Illinois at Urbana-Champaign Grant RB21012. We acknowledge Dr. Ymkje Anna de Vries and Prof. Marcus Munafò for making their dataset available under the Non-Commercial Government Licence v2.0 and Zhonghe Wan for her assistance in statistical calculations.

References

Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2017). Selective citation in the literature on swimming in chlorinated water and childhood asthma: A network analysis. *Research Integrity and Peer Review*, *2*(1). http:/doi.org/10.1186/s41073-017-0041-z

Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2019). Selective citation in the literature on the hygiene hypothesis: A citation analysis on the association between infections and rhinitis. *BMJ Open*, *9*(2), e026518. http:/doi.org/10.1186/s41073-017-0041-z

Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. *BMJ*, *339*, b2680. https://doi.org/10.1136/bmj.b2680

Greenberg, S. A. (2011). Understanding belief using citation networks: Citation networks. *Journal of Evaluation in Clinical Practice*, *17*(2), 389–393. http://doi.org/10.1111/j.1365-2753.2011.01646.x

Hsiao, T.-K., Fu, Y., & Schneider, J. (2020). Visualizing evidence-based disagreement over time: The landscape of a public health controversy 2002-2014. *Proceedings of the Association for Information Science and Technology* (Vol. 57, p. e315). https://doi.org/10.1002/pra2.315

Pariser, Eli. (2011). *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. New York: Penguin Press.

Schrag, M., Mueller, C., Oyoyo, U., Smith, M. A., & Kirsch, W. M. (2011). Iron, zinc and copper in the Alzheimer's disease brain: A quantitative meta-analysis. Some insight on the influence of citation bias on scientific opinion. *Progress in Neurobiology*, *94*(3), 296–306. https://doi.org/10.1016/j.pneurobio.2011.05.001

Sen, S., & Gan, S. K. (1983). A mathematical extension of the idea of bibliographic coupling and its applications. *Annals of Library Science and Documentation*, *30*, 78–82.

Shen, S., Zhu, D., Rousseau, R., Su, X., & Wang, D. (2019). A refined method for computing bibliographic coupling strengths. *Journal of Informetrics*, *13*(2), 605–615. https://doi.org/10.1016/j.joi.2019.01.012

Trinquart, L., Johns, D. M., & Galea, S. (2016). Why do we think we know what we know? A metaknowledge analysis of the salt controversy. *International Journal of Epidemiology*, *45*(1), 251–260. https://doi.org/10.1093/ije/dyv184

Urlings, M. J. E., Duyx, B., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2019a). Selective citation in scientific literature on the human health effects of bisphenol A. *Research Integrity and Peer Review*, *4*(1), 6. https://doi.org/10.1186/s41073-019-0065-7

Urlings, M. J. E., Duyx, B., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. A. (2019b). Citation bias in the literature on dietary trans fatty acids and serum cholesterol. *Journal of Clinical Epidemiology*, *106*, 88–97. https://doi.org/10.1016/j.jclinepi.2018.10.008

de Vries, Y. A., Roest, A. M., Franzen, M., Munafò, M. R., & Bastiaansen, J. A. (2016). Citation bias and selective focus on positive findings in the literature on the serotonin transporter gene (5-HTTLPR), life stress and depression. *Psychological Medicine*, *46*(14), 2971–2979. https://doi.org/10.1017/S0033291716000805

de Vries, Ymkje Anna, & Munafò, M. (2016). [Dataset] Citation bias and selective focus on positive findings in the literature on 5-HTTLPR, life stress, and depression. University of Bristol. Retrieved January 28, 2021, http://doi.org/10.5523/BRIS.Z7JCONXFBMDR1JJ3T0W4K1HWN

Two-Dimensional Mapping of University Profiles in Research.

Joel Emanuel Fuchs[1] and Thomas Heinze[2]

[1] *jfuchs@uni-wuppertal.de*
University of Wuppertal, Institute of Sociology, Gauss-Str 20, D-42119 Wuppertal (Germany)

[2] *theinze@uni-wuppertal.de*
University of Wuppertal, Institute of Sociology & Interdisciplinary Center for Science and Technology Studies (IZWT), Gauss-Str 20, D-42119 Wuppertal (Germany)

Abstract

This paper presents a two-dimensional graphical mapping for institutional profiles of higher education entities (such as colleges and universities) in research, teaching, technology transfer, or internationalization. For the sake of simplicity, our illustrations focus on research profiles only. The new graphs embrace both the organizational field level and the organizational level. We illustrate how the new graphical representation improves existing ones, using examples from the German public university system.

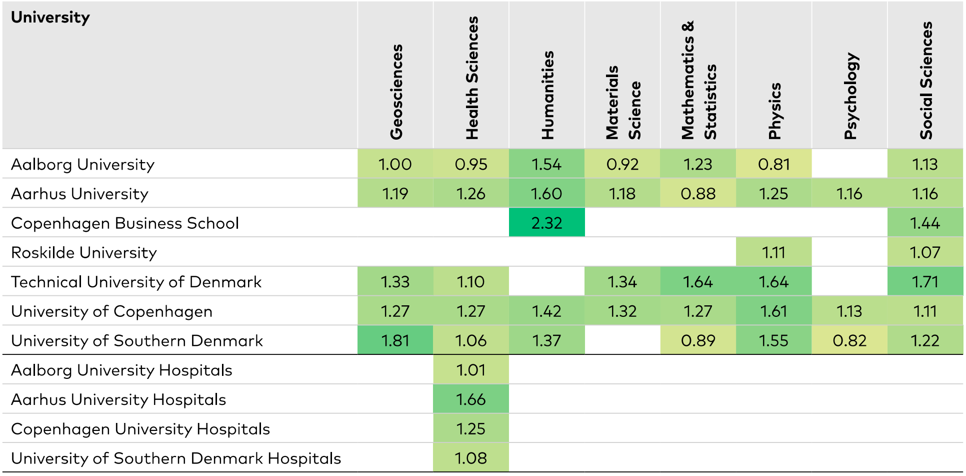
Introduction

Mapping profiles in higher education, especially with regard to research and teaching, has proliferated in recent years. Such mapping is done in the context of an increased attention to large-scale visualization of science and technology (Borner, Bueckle, & Ginda, 2019; Fortunato et al., 2018). One focus in this literature have been comparisons at the system level (van Vught, 2009). An example is Huisman, Lepori, Seeber, Frölich, and Scordato (2015) who classify 24 national higher education systems in Europe with regard to their degree of horizontal differentiation in research, teaching, technology transfer, and internationalization. Another example is Harzing and Giroud (2014) who identify top-3/bottom-3 research areas in 34 national higher education systems using the *Revealed Comparative Advantage (RCA)* measure. At the country level, Teixeira, Rocha, Biscaia, and Cardoso (2012) found for Portugal that private higher education entities have research profiles complementary to those of public entities, using the *Relative Specialization Index (RSI)*.

A second focus in the literature is on organizational fields in higher education (Brint et al., 2011; Ruef & Nag, 2015). A well-known comprehensive mapping of bibliometric profiles has been done for the Nordic universities (Piro et al. 2017; 2014). Here, the *RSI* is used for the comparison of field-related publication and citation percentages with respective global field percentages, highlighting those colleges and universities with either below-average, average, or above-average contributions. In addition, Bonaccorsi, Colombo, Guerini, and Rossi-Lamastra (2013), by means of the *Activity Index (AI)*, show for Italy that universities specialized in applied fields and engineering have a positive impact on start-ups in their region, especially in the service industries. In contrast, universities with a profile in basic science fields are related to a greater number of start-ups in manufacturing. More recently, teaching and research profiles of public universities in Germany were mapped, using both the *RSI* and a modified version of it, the *RESP* (details below), with longitudinal data on scientific staff, funding, bibliometric indicators, and student enrolment (Heinze, Tunger, Fuchs, Jappe, & Eberhardt, 2019).

Mapping institutional profiles typically involves graphical representations, often *heatmaps*, that are meant to display degrees of specialization. Consider the example in Figure 1. Such heatmaps depict specializations with a color-spectrum ranging from yellow to green (Piro et al. 2017; 2014). Although such heatmaps show whether a university has either a below-average or above-average profile in a research field, they do not display the size of the respective field in the university under consideration. In other words: although a university might be specialized in a particular field, such as Aarhus University in “humanities” (Figure 1), the heatmap does not provide information with respect to the size of the humanities compared to the 7 other fields in

Aarhus. Of course, this information could be provided in a separate table or graph, but this would make comparisons of large numbers of higher education entities rather complex.



University	Geosciences	Health Sciences	Humanities	Materials Science	Mathematics & Statistics	Physics	Psychology	Social Sciences
Aalborg University	1.00	0.95	1.54	0.92	1.23	0.81		1.13
Aarhus University	1.19	1.26	1.60	1.18	0.88	1.25	1.16	1.16
Copenhagen Business School			2.32					1.44
Roskilde University						1.11		1.07
Technical University of Denmark	1.33	1.10		1.34	1.64	1.64		1.71
University of Copenhagen	1.27	1.27	1.42	1.32	1.27	1.61	1.13	1.11
University of Southern Denmark	1.81	1.06	1.37		0.89	1.55	0.82	1.22
Aalborg University Hospitals		1.01						
Aarhus University Hospitals		1.66						
Copenhagen University Hospitals		1.25						
University of Southern Denmark Hospitals		1.08						

Figure 1. One-dimensional heatmap, using a yellow-green color spectrum
Source: RSI scores for citation rates (2011-2014), Piro et al. (2017, p. 82).

This paper provides a new graphical representation that depicts both the specialization of a given higher education entity in a particular academic discipline (relative to all other relevant higher education entities) and its size within that entity. This new representation displays two types of information simultaneously. First, it shows whether a college or university either has a below-average, average, or above-average profile in a given research field. Second, it shows whether the research field is big, mid-size, or small compared to all other fields within that higher education entity. We provide examples from the German public university system, illustrating how the new graphical representation improves the existing one.

Data and Method

Our analysis builds on a dataset of 68 public universities in Germany, with information on scientific staff, basic funding, grant funding, publications, citations, and enrolment available for the years 1992-2018 (for details see Heinze et al., 2019). Based on these data, we calculated and then visualized institutional profiles, using both the RSI and the RESP. Results are available on the following website: https://fachprofile.uni-wuppertal.de/en.html. Both RSI and RESP are based on the aforementioned *Activity Index* (Narin, Carpenter, & Woolf, 1987). The *AI* captures the extent to which certain entities are specialized in certain activities (Formula 1). *AI* values lower 1.0 indicate a negative specialization (below-average scores), *AI* values greater 1.0 indicate a positive specialization (above-average).

Formula 1: General formula of the Activity Index (AI)

$$AI_{ij} := \frac{N_{ij}/\sum_i N_{ij}}{\sum_j N_{ij}/\sum_{ij} N_{ij}}$$

The AI's value range of [0.0,+∞] lacks an upper limit. Therefore, interpreting the RSI is easier than the AI due to its symmetrical value range of [−1.0,+1.0]: values lower 0.0 indicate negative specialization; values greater 0.0 indicate positive specialization. We use a modified version of the RSI that was introduced by (Grupp 1994; 1998). Its value range is [-100.0, +100.0] with an expected value of zero (Formula 2). We call this index RESP (for "Index of Relative Specialization"). It is based on the *hyperbolic tangent.* Consequently, its curve is steeper and reaches the upper limits of its value range more quickly than RSI.

Formula 2: Relative Specialization (RESP)

$$\mathrm{RESP} := 100\,\frac{\mathrm{AI}^2 - 1}{\mathrm{AI}^2 + 1}$$

Note: The subindices i and j of the AI are omitted for the sake of simplicity.

Consider Figure 2 as an example that displays RESP scores for Web of Science publications of the University of Aachen from 1995-2018. At least three results can be inferred from the heat-map. *First*, Aachen has an above-average publication profile in psychology in all years. *Second*, Aachen's publication profile in economics has shifted from above-average to below-average. *Third*, Aachen's publication profile in chemistry has changed from below-average to average. As mentioned above, Figure 2 does not display the size of disciplines at the University of Aachen: the color bars of psychology, economics, chemistry are all of the same size. However, in reality, the three disciplines have very different publication outputs – and this is true also for other indicators, such as professorial staff or funding. As shown in Figure 3, Aachen's largest discipline in terms of publications is physics/astronomy ("Physik, Astronomie"), its bar has maximum size. All other disciplines are measured relative to it. For example, chemistry is smaller than physics/ astronomy but larger than psychology or economics. In other words: specialization in psychology requires a much smaller number of publications compared to physics/astronomy. As Figure 3 shows, physics/astronomy has most publications in Aachen, but their number is not as high as to constitute a particularly strong specialization in this discipline: its scores are yellow, meaning Aachen's number of physics publications are on average, compared to all other (here: technical) universities with publications in this discipline. The different bar sizes of disciplines in Figure 3 are based on Formula 3 for a given year. A university's largest discipline has bar size =1, because the enumerator equals the denominator, whereas all other disciplines have a bar size relative to the largest discipline. In Aachen, as mentioned, the largest discipline with regard to publications is physics/astronomy. The three other above-mentioned disciplines have the following bar sizes [min, max] between 1995 and 2018: chemistry [0,367; 0,830]; psychology [0,024; 0,140]; economics [0,010; 0,032]. All annual values are placed in the middle of the bar, and adjacent bars are connected using a natural cubic spline interpolation which has the effect of smoothening the bar graph.

Formula 3: Bar size calculation in Figures 3-5

$$\text{Bar size} := \frac{\text{Publications of University i in Discipline j}}{max_j\ \text{(Publications of University i in Discipline j)}}$$

Results

Figure 3 constitutes an improvement compared to Figure 2 because it allows to simultaneously evaluate both disciplinary specialization (measured at the organizational field level) and size of disciplines (measured at the organizational level). What are possible conclusions? *First*, Figure 3 depicts disciplines that have been for a long time below-average (geosciences) or have become so (economics, mathematics). *Second*, these three disciplines are marginal in terms of overall publication output; here four disciplines dominate: physics/astronomy, chemistry, mechanical and process engineering, and biology. *Third*, Aachen's increasing specialization in chemistry has been the result of a considerable growth in publications, relative to the discipline with most publications over time (physics/astronomy).

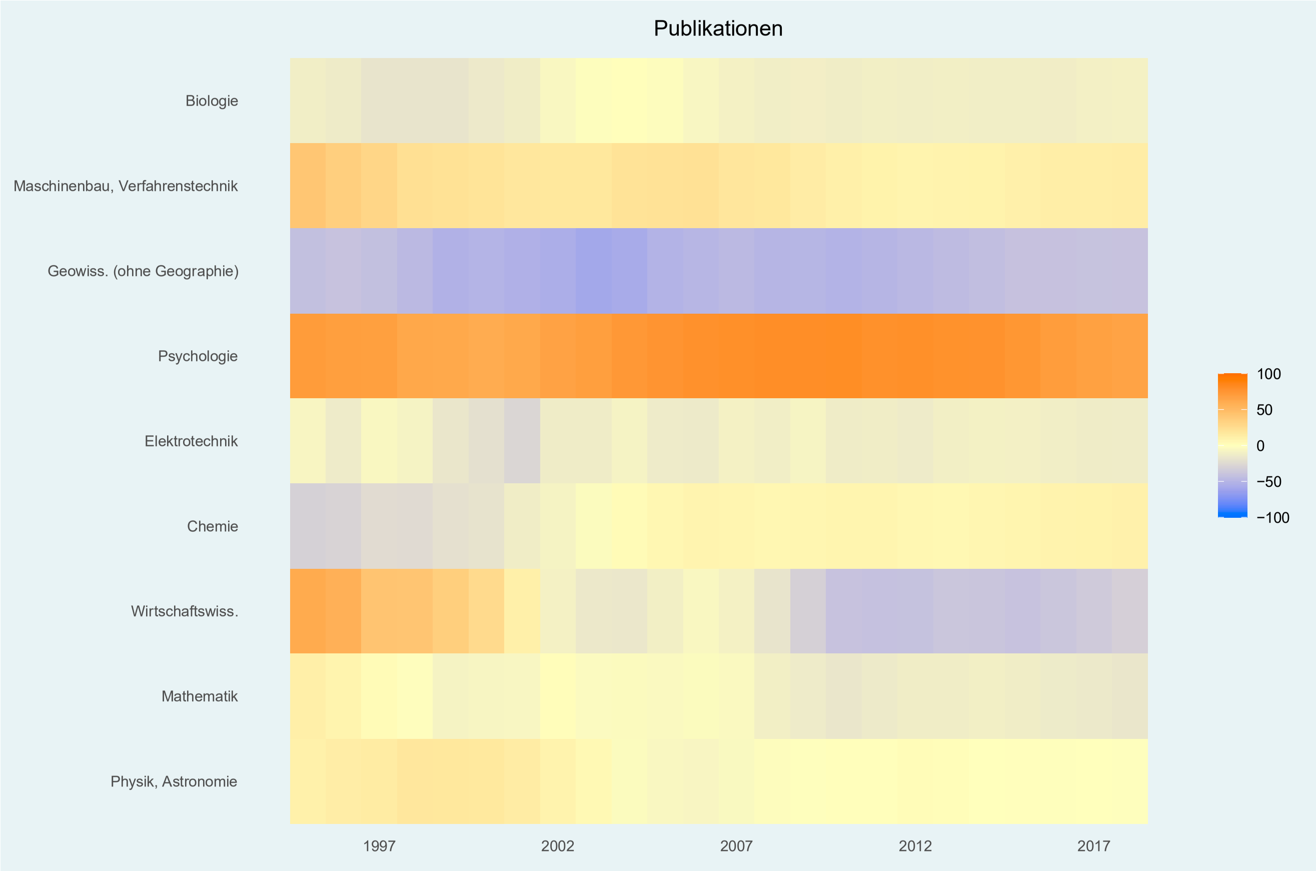


Figure 2. One-dimensional heatmap, Web of Science publications (Univ. of Aachen)
Source: RESP scores for WoS publications, available at: https://fachprofile.uni-wuppertal.de/en.html.

This interpretation for Aachen's publication profile can be squared with other indicators, such as scientific staff. Consider Figure 4 that captures Aachen's development of professorial staff. The overall picture looks quite different from that of publications in that almost all disciplines have grown relative to the discipline with most professors (mechanical/ process engineering). Two conclusions seem warranted regarding the comparison of Figures 3 and 4. *First*, in some disciplines, we observe double growth in input (professorial staff) and output (publication output) that leads to a stronger disciplinary profile (e.g., chemistry, physics/astronomy). *Second*, in other disciplines, input and output are decoupled, and there seems to be no coherent institutional profile (e.g., economics, geosciences, mathematics).

Another functionality of the new graphical representation is that it documents the connection between changes both in funding structure and research profiles. For example, consider the University of Wuppertal where grant funding has considerably changed since the late 1990s (Figure 5). In particular, the growth of externally funded research projects in mathematics moves the university's research profile in this discipline from below-average (blue) to above-average (orange) within a few years. Note that mathematics is deeply orange although its grant funding is only a fraction of the largest discipline (electrical engineering). In contrast, chemistry takes almost the opposite route: within ten years only, the university switches its research profile in this discipline from above-average to below-average. Here, the decrease in grant funding occurs simultaneously: both on the organizational field level (other universities) and on the organizational level (grant funding, Wuppertal).

Discussion

We introduce a two-dimensional graphical mapping, with emphasis on research profiles of public universities. The new graphical representation can be applied to other dimensions, such as teaching, technology transfer, or internationalization as well. The key difference compared to existing *heatmaps* is that our new graphs capture comparisons on both the organizational field level (here: other universities) and the organizational level (university). In this way, we make a first step in better understanding the interplay between both levels. We are aware that our contribution is descriptive, and that further statistical analyses regarding the two levels and their function for both building and maintaining institutional profiles are necessary.

Regarding the graphical representation in Fig. 3–5, one might wonder about other technical possibilities, such as Sankey diagrams, stack area charts, or stream graphs (Wickham, 2016). Although it would be possible, for example, to stack the heat bars, and thus arrive at stack area charts, we decided not to do this. First, our prime motivation is to improve existing (and widely used) heat maps (as discussed above). Therefore, we used existing heat maps available for German public universities (https://fachprofile.uni-wuppertal.de). By adding one additional slice of information (here: the share of a field in comparison to all fields at one university), we wanted the new graphs to be as similar as possible to the existing ones, resulting in variable heights of the heat bars without stacking. Second, when using stack area charts or stream graphs, every research field would get their own color (so fields can be identified). But this is not our aim. In our graphs, colors represent RESP values. Consequently, different fields are displayed by the same color. In contrast, stalking the field would lead to a much more complex color spectrum in the graph, and thus counteract our main purpose: the production of intuitively understandable graphical representations that capture both the field and organizational levels.

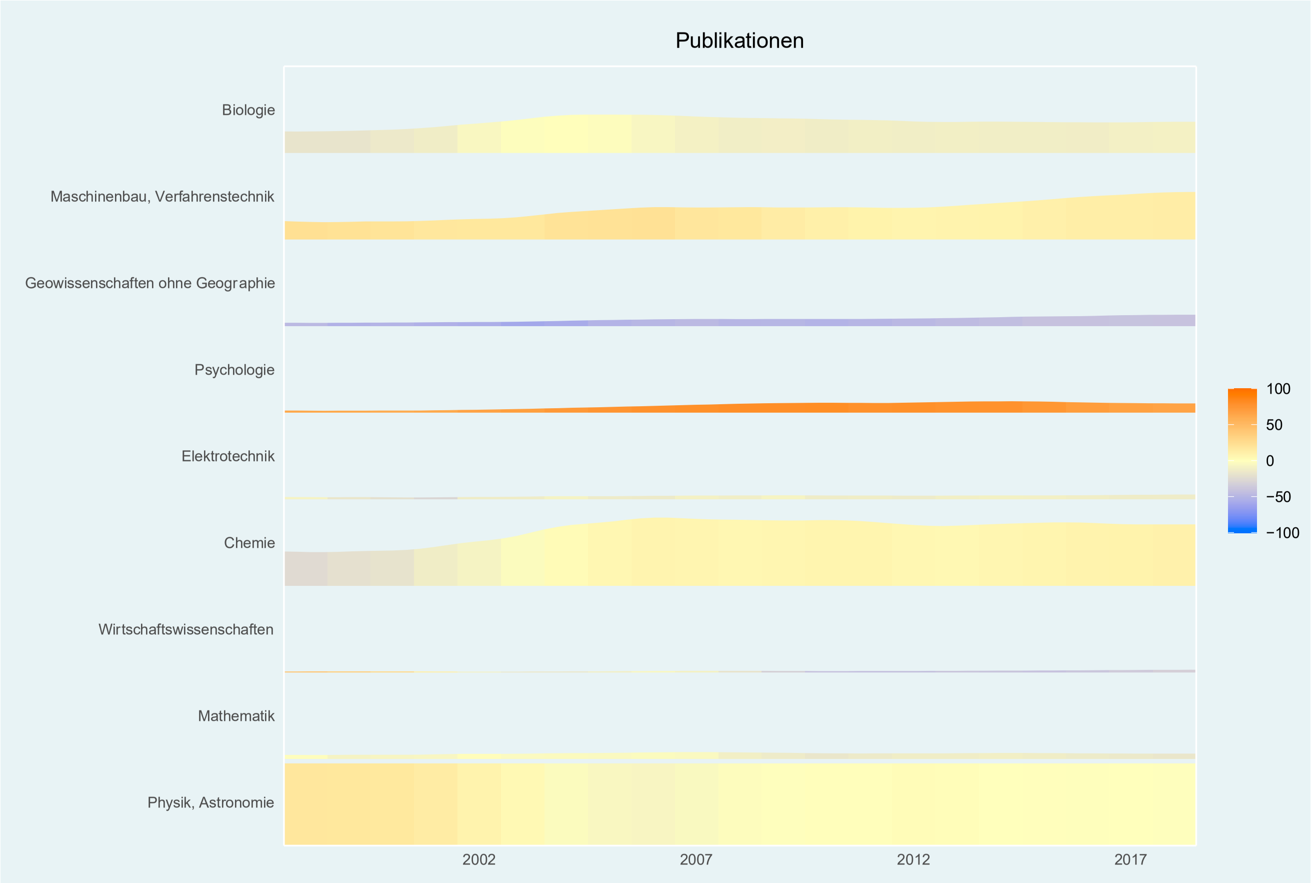


Figure 3. Two-dimensional heatmap, Web of Science publications (Univ. of Aachen)
Source: RESP scores for WoS publications plus respective size of disciplines.

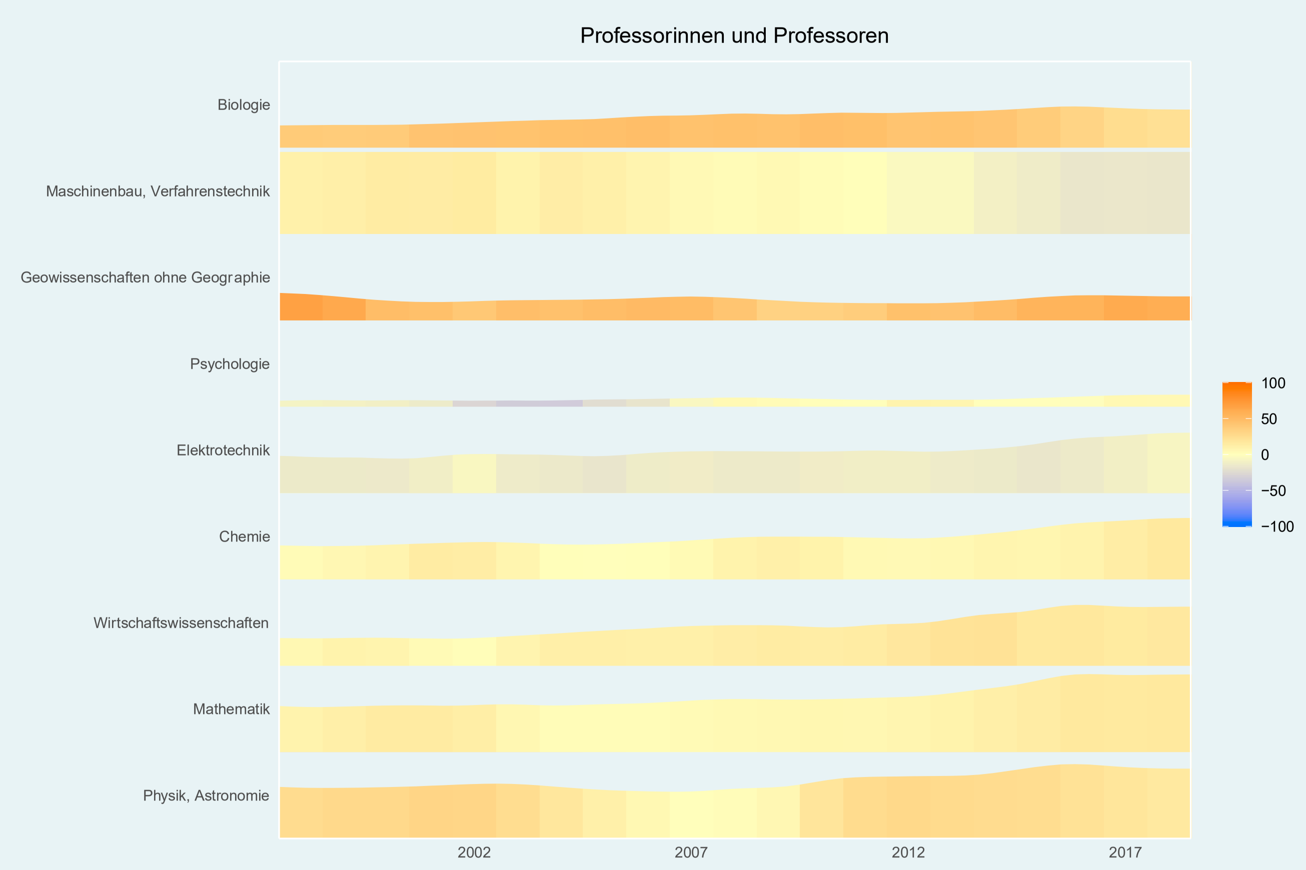


Figure 4. Two-dimensional heatmap, professorial staff (Univ. of Aachen)
Source: RESP scores for professorial staff plus respective size of disciplines.

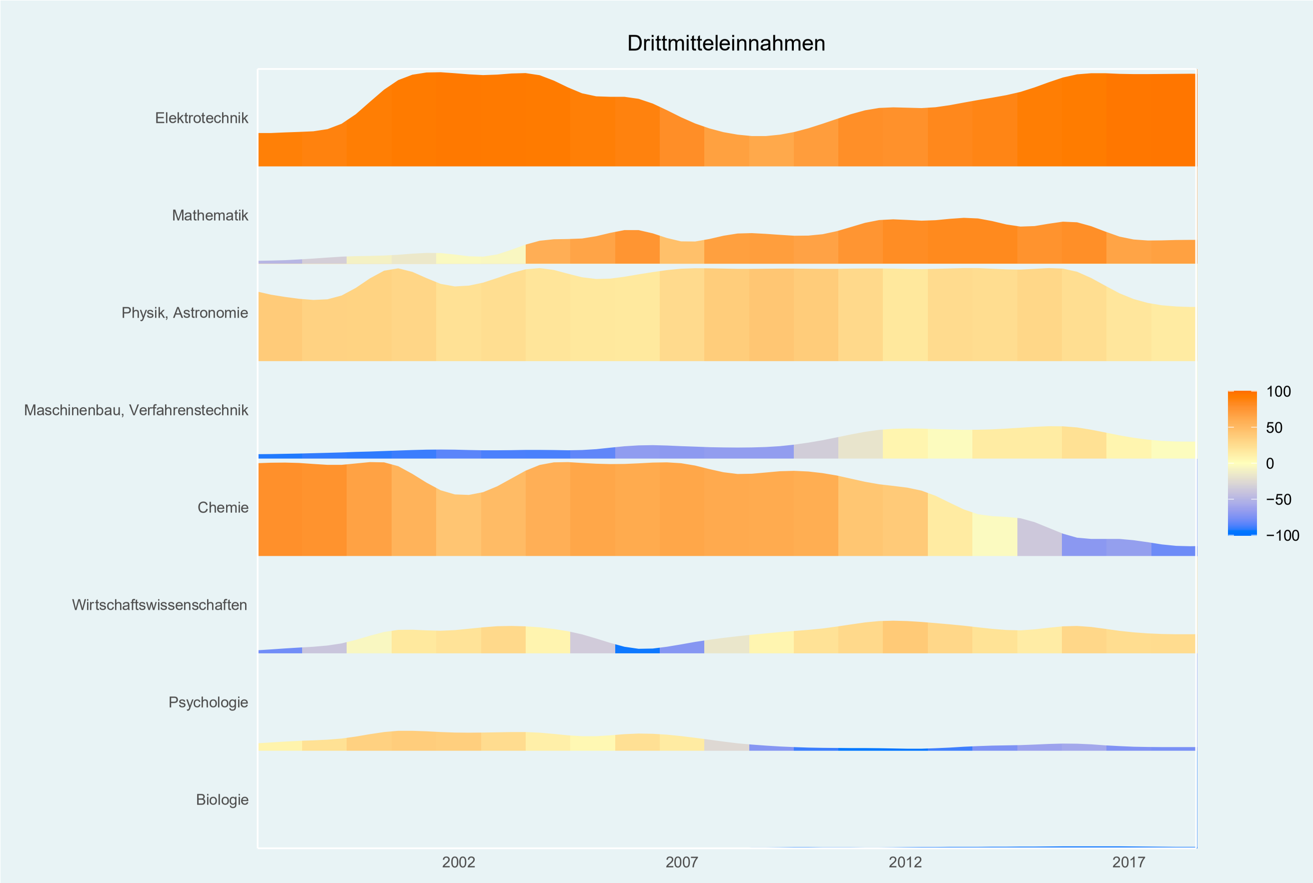


Figure 5. Two-dimensional heatmap, grant funding (Univ. of Wuppertal)
Source: RESP scores for grant funding plus respective size of disciplines.

Notes

All graphs can be found and used under the CC-BY-NC-ND-4.0 international license at https://fachprofile.uni-wuppertal.de/conferences/issi2021.html. Analysis was conducted in R (R Core Team, 2020) with data.table (Dowle & Srinivasan, 2019) and figures were produced using ggplot2 (Wickham, 2016). For smoothing the heat bars, the packages stats (R Core Team, 2020) and splines (ibid.) were used. The pseudo code for smoothing is (as geometric part of the ggplot command): stat_smooth(method = 'glm', method.args = list(family = gaussian), formula = y ~ splines::ns(x,df = years - 1), se = FALSE, geom = "ribbon", span = 1).

Acknowledgments

We thank Dirk Tunger and Paul Eberhardt for their support in data collection and curation.

References

Bonaccorsi, A., Colombo, M. G., Guerini, M., & Rossi-Lamastra, C. (2013). University specialization and new firm creation across industries. *Small Business Economics, 41*, 837-863.

Grupp, H. (1994). The measurement of technical performance of innovations by technometrics and its impact on established technology indicators. *Research Policy, 23*, 175-193.

Grupp, H. (1998). Measurement with patent and bibliometric indicators. In H. Grupp (Ed.), *Foundations of the economics of innovation. Theory, Measurement, Practice* (pp. 141-188). Cheltenham: Edward Elgar.

Harzing, A.-W., & Giroud, A. (2014). The competitive advantage of nations. An application to academia. *Journal of Informetrics, 8*, 29-42.

Heinze, T., Tunger, D., Fuchs, J. E., Jappe, A., & Eberhardt, P. (2019). *Research and teaching profiles of public universities in Germany. A mapping of selected fields*. Wuppertal: BUW.

Huisman, J., Lepori, B., Seeber, M., Frölich, N., & Scordato, L. (2015). Measuring institutional diversity across higher education systems. *Research Evaluation, 24*, 369-279.

Narin, F., Carpenter, M. P., & Woolf, P. (1987). Technological assessments based on patents and patent citations. In H. Grupp (Ed.), *Problems of measuring technological change* (pp. 107-119). Köln: TÜV Rheinland.

Piro, F. N., Aldberg, H., Aksnes, D. W., Staffan, K., Leino, Y., Nuutinen, A., . . . Sivertsen, G. (2017). *Comparing research at nordic higher education institutions using bibliometric indicators covering the years 1999-2014. Policy Paper 4/2017*. Oslo: NIFU.

Piro, F. N., Aldberg, H., Finnbjörnsson, Þ., Gunnarsdottir, O., Karlsson, S., Skytte Larsen, K., . . . Sivertsen, G. (2014). *Comparing Research at Nordic Universities using Bibliometric Indicators – Second report, covering the years 2000-2012. Policy Paper 2/2014*. Oslo: NordForsk.

Teixeira, P. N., Rocha, V., Biscaia, R., & Cardoso, M. F. (2012). Competition and diversity in higher education: an empirical approach to specialization patterns of Portuguese institutions. *Higher Education, 63*(3), 337-352.

van Vught, F. A. (Ed.) (2009). *Mapping the Higher Education Landscape. Towards a European Classification of Higher Education*. Dordrecht: Kluwer.

Bonaccorsi, A., Colombo, M. G., Guerini, M., & Rossi-Lamastra, C. (2013). University specialization and new firm creation across industries. *Small Business Economics, 41*, 837-863.

Borner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences of the United States of America, 116*(6), 1857-1864. doi:10.1073/pnas.1807180116

Brint, S., Proctor, K., Hanneman, R. A., Mulligan, K., Rotondi, M. B., & Murphy, S. P. (2011). Who are the early adopters of new academic fields? Comparing four perspectives on the institutionalization of degree granting programs in US four-year colleges and Universities, 1970-2005. *Higher Education, 61*(5), 563-585. doi:10.1007/s10734-010-9349-z

Dowle, M., & Srinivasan, A. (2019). data.table: Extension of data.frame. R package version 1.12.8. Retrieved from https://CRAN.R-project.org/package=data.table

Fortunato, S., Bergstrom, C. T., Borner, K., Evans, J. A., Helbing, D., Milojevic, S., . . . Barabasi, A. L. (2018). Science of science. *Science, 359*(6379). doi:10.1126/science.aao0185

Grupp, H. (1994). The measurement of technical performance of innovations by technometrics and its impact on established technology indicators. *Research Policy, 23*, 175-193.

Grupp, H. (1998). Measurement with patent and bibliometric indicators. In H. Grupp (Ed.), *Foundations of the economics of innovation. Theory, Measurement, Practice* (pp. 141-188). Cheltenham: Edward Elgar.

Harzing, A.-W., & Giroud, A. (2014). The competitive advantage of nations. An application to academia. *Journal of Informetrics, 8*, 29-42.

Heinze, T., Tunger, D., Fuchs, J. E., Jappe, A., & Eberhardt, P. (2019). *Fachwissenschaftliche Forschungsprofile staatlicher Universitäten in Deutschland. Eine Kartierung ausgewählter Fächer*. Wuppertal: BUW.

Huisman, J., Lepori, B., Seeber, M., Frölich, N., & Scordato, L. (2015). Measuring institutional diversity across higher education systems. *Research Evaluation, 24*, 369-279.

Narin, F., Carpenter, M. P., & Woolf, P. (1987). Technological assessments based on patents and patent citations. In H. Grupp (Ed.), *Problems of measuring technological change* (pp. 107-119). Köln: TÜV Rheinland.

Piro, F. N., Aldberg, H., Aksnes, D. W., Staffan, K., Leino, Y., Nuutinen, A., . . . Sivertsen, G. (2017). *Comparing research at nordic higher education institutions using bibliometric indicators covering the years 1999-2014. Policy Paper 4/2017*. Oslo: NIFU.

Piro, F. N., Aldberg, H., Finnbjörnsson, Þ., Gunnarsdottir, O., Karlsson, S., Skytte Larsen, K., . . . Sivertsen, G. (2014). *Comparing Research at Nordic Universities using Bibliometric Indicators – Second report, covering the years 2000-2012. Policy Paper 2/2014*. Oslo: NordForsk.

Ruef, M., & Nag, M. (2015). *The classification of organizational forms: Theory and application to the field of higher education*. Stanford: Stanford University Press.

Team, R. C. (2020). *R: A language and environment for statistical computing.* . Retrieved from Vienna: https://www.R-project.org/.

Teixeira, P. N., Rocha, V., Biscaia, R., & Cardoso, M. F. (2012). Competition and diversity in higher education: an empirical approach to specialization patterns of Portuguese institutions. *Higher Education, 63*(3), 337-352.

van Vught, F. A. (Ed.) (2009). *Mapping the Higher Education Landscape. Towards a European Classification of Higher Education*. Dordrecht: Kluwer.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer.

Academic Leadership and University Performance: Do Russian Universities Improve after They Are Headed by Top Researchers?

Daria Gerashchenko

dgeraschenko@eu.spb.ru
European University at St Petersburg (Russian Federation)

Abstract

Studies that have examined organizations' productivity and their leaders have found a rather weak relationship between a leader's personal characteristics and organizational output. In this study, I take a theoretical approach to quantify the effects of top leadership on university research performance. I assume that top leaders may influence university research productivity, but this influence should be visible. I theorize two types of university leader: the "strategic manager" who seeks to reallocate resources to make it of use for the whole university and the "politician" who reallocates resources to benefit certain research areas, especially the one they specialize in. Using an extensive sample of Russian universities, I demonstrate that while there is no observable relationship between overall university research productivity and the university leader's academic excellence, there is a positive influence by university leader's academic productivity on the research performance of his/her specific research field.

Introduction

How important is leadership for how organizations perform? Organizational scientists in general believe that institutional leaders are in a strategic position to influence the implementation of organizational goals (Thornton and Ocasio 1999; Pfeffer 1981; Yielder and Codling 2004; Kraatz and Moore 2002; Henkel 2002; Hambrick and Mason 1984; Fligstein 1987). Good leaders make organizations more consistent and stable while also enabling them to achieve certain goals and be more effective. Leaders function as key actors that provide a link between the organization and its environment based on their reading of the environment and they use resources and power to initiate new actions (Fligstein 1987). However, in professional organizations, leaders' potential to influence institutional courses of action can be restricted (Birnbaum 1989; Cohen and March 1974). Recent studies on academic leadership demonstrate that the question is not whether there is a link between leadership and organizational performance, but what type of leaders or under what conditions leaders are in a position to influence university performance. This study, based on extensive data from Russian universities, joins other attempts (Goodall 2005; 2009; Goodall, McDowell, and Singell 2017) to analyze the impact of academic leaders on organizational performance.

Who is best prepared to govern academic institutions in the most effective way? Conventional assumptions suggest that the best leader combines academic background and managerial experience (Etzioni 1959). However, since the recent managerial revolution, traditional academic experience has begun to be perceived as less important than the managerial background necessary to turn universities into effective organizations (Breakwell and Tytherleigh 2008; Engwall 2014; Musselin 2007). The empirical research suggests that the academic excellence of university leaders is statistically significant for the research quality of the university they lead (Goodall 2005; 2009; Goodall, McDowell, and Singell 2017). However, Breakwell and Tytherleigh (2010) studied a broad sample of UK universities and found that variability in the performance of a university is explained by non-leadership factors. In this study, I raise the same question and ask whether the research performance of an academic leader influences the institutional performance of the organization they lead. Contrary to previous empirical studies, I focus on the leaders' research field specialization. The results demonstrate that a leader's academic performance effect is more profound in the areas of research in which the leader specializes. The findings thus support the argument for the importance of expert specialization.

This paper looks specifically into the relationship between a university leader's individual academic performance and the university's research performance. Research question is thus: "How does university productivity change if the university leader is a top scholar?". In contrast to the largely untheoretical approaches of previous studies exploring the same question, I consider two major theories to help explain the existence—or its absence—of a relationship between a university research performance and its leader's academic achievements. One is the managerial model which assumes that university leaders act as a fair resource reallocator to boost the university performance in general because s/he is deeply invested in the university and plans either to work in the same position at other institutions or to retire as university leader. The second is the political model which assumes that a university leader will favor the department of his/her research specialization to boost its performance by providing scarce resources especially if s/he was previously affiliated with a particular department or plans to work there after the term as leader is over.

This study doesn't seek to forecast future performance, but rather to establish correlation between the variables of interest. For this study, I use Russian Index of Scientific Citation (hereafter RISC) as main data source. I collected individual research performance and university research performance information for 915 universities and 949 rectors. First, I analyze whether there is a correlation between university performance in general and whether the university's rector is a leading scholar is his or her field. Second, I look at how research the various academic fields perform academically at the university and inquire whether the fact that the university leader is a specialist in a specific field is statistically related to it. The models are tested using panel data.

Theoretical framework

The question of the president's influence on university performance was raised in a pioneering paper by Amanda Goodall (2006). In this paper, she analyzed top research universities and found a positive correlation between a leader's academic achievements—measured by citation count—and the research university's international ranking. Breakwell and Tytherleigh's (2010) study of British universities' KPIs and leader's background, shows somewhat different results than Goodall's. They find that university leaders' characteristics as well as leadership change have no observable influence on university performance. These conflicting findings suggest that the effect of top leadership may not be necessarily visible for the university in general. So far, the possibility of rector having an effect on research area's academic performance hasn't been studied while rectors are closely attached to their background areas. Present study fills this gap by exploring the effect of top leadership on the research performance of the different research areas. I do this by positing two theoretical types of leaders. The first type relies on a New Public Management (hereafter NPM) studies according to which university leaders are strategic managers seeking to reallocate resources for the betterment of the whole university. The second type takes its bearing from the political model of organizational governance. This model argues that leaders are politicians who reallocate resources to benefit particular constituencies or research fields.

From a NPM perspective, a university leader should be someone who seeks to balance the competing forces within the university and avoid conflict while also effectively delivering for the betterment of the whole university. This logic sees university leaders as strategic managers setting the rules of the game for all university departments regardless of their own departmental specialization. This quality of benefiting all university departments equally is considered a crucial feature for successful university leaders following the principles of objectivity in decision-making (Julius, Baldridge, and Pfeffer 1999). Based on the mentioned above research literature on the effect of top researchers on the organizational performance (Goodall 2009; Goodall, McDowell, and Singell 2017), I suggest that an academic success of university leaders might have an influence on the research output of universities. This allows to formulate the first hypothesis.

H1a: There is a positive relationship between the academic success of a university leader and the future research performance of university in general.

I further propose to examine whether the academic success of new leadership is more strongly related to university performance for externally recruited leaders rather than from within. In contrast to leaders selected from within an organization, new leaders brought from the outside might be less bound by implicit contracts and internal coalitions (Cannella Jr and Lubatkin 1993; Staw 1980). In addition, outsider hires can be perceived as more capable than insiders to initiate and implement strategic changes (Zhang and Rajagopalan 2004; Helmich 1975; Cannella Jr and Lubatkin 1993). Thus, I propose:

H1b: A positive relationship between the academic success of a university leader and the future research performance of university in general will be found only for externally recruited leaders.

The second leadership type I consider is the "political model." This model was developed largely as a contrast to the bureaucratic model which assumed that formal authority is limited by power blocs and interest groups internal to the organization (Baldridge 1971). A number of studies argue that most organizations are of political in nature this sense and universities are not an exception (Salancik and Pfeffer 1974; Pfeffer and Moore 1980). Drawing on this type, I suggest that the power that a university leader possesses may come not only from his/her formal position, but also from intra-organization interest groups located at the subunit level (Pfeffer and Salancik 1974; Salancik and Pfeffer 1974). I expect university leader to act more as a politician deliberately reallocating resources in favor of certain university departments with which s/he is affiliated especially if the leader is an internal candidate. Those university leaders who had previously worked at the university and/or specialized in a specific research area are part of a tight intra-organizational network and are expected to favor these networks while in office. Based on this, I propose the second hypothesis.

H2a: There is a positive relationship between the academic success of a university leader and the future research performance of the field in which the leader specialized or previously worked.

H2b: The effect of the relationship between the academic success of a university leader and the future research performance of the field in which the leader specialized will be stronger for internally recruited leaders.

Data and methods

University academic productivity can be measured in different ways. For example, it can be measured as the university's position in international rankings (Goodall 2005) or in national ratings (Goodall 2009), as a share of the total country-wide weighted publications (Goodall, McDowell, and Singell 2017). For this study we use the Comprehensive Publishing Performance Score (CPPS) as a measure of university performance and a key dependent variable. The CPPS (*Kompleksnyj ball publikacionnoj rezul'tativnosti*) is a numerical measure of research productivity introduced by the Russian Ministry of Education and Science and is based on university research performance reports, WoS data, and Ministry data. RISC uses the methodology of CPPS calculation and publishes it. As RISC is a national bibliometric database and covers a large amount of bibliometric data (Sīle et al. 2018) across the Russian academy in comparison to WoS and Scopus, I consider CPPS to be a more reliable measure of university research productivity in Russia. RISC provides the opportunity to count CPPS separately for research fields (life sciences, social sciences, humanities, etc.). While it is possible to use administrative data about the research performance of Russian universities as others have demonstrated (Dyachenko and Mironenko 2019), the bibliometric data provided by RISC core is more reliable as it is independently collected.

Although I acknowledge some difficulties that may arise from using citations as an individual research output measure (Goodall 2005, 390–91), citations are important for understanding academic achievements. A number of studies use citations to assess academic output and productivity (Bornmann et al. 2008; Bayers 2005; Taylor 2011; Goodall 2005; Goodall, McDowell, and Singell 2017; Goodall 2009). I consider two variables that determine a rector's academic achievement measure, i.e., a binary variable *top 10% scientist* and a continuous *p-score*. The *p-score* as a key measure of a rector's academic achievement was calculated using Goodall's (2005) approach. Moreover, since disciplines differ in their patterns of academic citations, I calculated citation thresholds for each discipline using RISC lists to create a list of the *top 10%* best researchers in each field. A rector's p-score thus equals the rector's citation number divided by a threshold. While a binary *top rector* variable equals 1 if the rector is among the top 10% most cited academics in his/her research area according to RISC lists for core RISC and RISC in general. As both measures showed virtually the same results in the preliminary phase of the analysis, I choose the p-score as a more sensitive measure of change in individual rectors' performances. In addition, I gathered biographical data of university leaders from official university websites and government portals as well as the leaders' personal web pages.

Table 1. Summary statistics for key variables (restricted sample).

VARIABLES	LABELS	N	mean	sd	min	max	median
year_founded	Year university was founded in	705	1959	45.05	1712	2016	1968
sex	Rector's sex (1-male, 0 - female)	705	0.79	0.41	0	1	1
year_pub	Year of rector's first publication	705	1993	14.85	1955	2017	1996
project5100	University-participant of the project (1-yes,0-no)	705	0.03	0.17	0	1	0
top10_yes	Top rector (1-yes, 0-no)	705	0.34	0.47	0	1	0
N_authors	Number of authors at the university	705	1005.6	1887.51	0	30293	455
cat_rector	Rector's specialization (area of expertise), encoded	705	29.1	13.3	1	49	25
term	Rector's term length in years	571	12.3	7.15	4	41	10
category_uni_encoded	University type, encoded	705	26.8	11.41	1	42	30
internal_rector	Rector is an internal candidate (1-yes,0-no)	556	0.68	0.47	0	1	1
external_rector	Rector is an external candidate (1-yes,0-no)	556	0.32	0.47	0	1	0
p-score	Rector's p-score 2010-2019	7.830	1.41	7.20	0.00	148.90	0.27
year	Year of observation	7.050	2014.5	2.872465	2010	2019	2014.5
CPPSALL	Overall CPPS 2010-2019	7.024	185.4	589.29	0.00	14310.7	54.8

Note: The last three variables are variables in a panel format.

Control variables are essential for reliable results. Control choice has been driven by several factors. It is necessary to account for individual and institutional factors that may have an influence on university research productivity. For instance, as some universities are the participants of the Russian Excellence Initiative (project 5-100), they may have better chances at scoring higher in research productivity variables than non-participants due to additional resources provided by the state. At the same time, younger universities may be less responsive to a rector's research activity in contrast to older institutions as others have demonstrated

(Engwall 2014). Importantly, university size may also contribute significantly to research productivity because in contrast to smaller universities, bigger ones have more resources to hire qualified employees who produce more academic output (Breakwell and Tytherleigh 2010; Goodall 2009).

Because not all rectors and universities have profiles in RISC, the initial sample that contained 951 universities (both public and private) and 949 rectors needed to be cut in size. The sample was also restricted to organizations in which the rector has been in office for no less than four years. As rectors' term of office in Russia is set at five years, I consider four years to be the minimum number of years needed to observe the effect of a leader on university performance. Since the main dependent variable covers the period from 2010 to 2019 (no prior data is available), I assumed that rectors who retired prior to 2007 could not affect the outcome variable. The result was 705 universities and 783 rectors (Table 1).

Results

First model results: overall university performance and top leadership

The first model seeks to test the hypothesis that there is a statistical relationship between Russian rectors' academic achievement and overall university research performance. The outcome variable is *overall CPPS 2010-2019* variable; the key independent variable is a continuous *p-score 2010-2019.* The data consists of a panel of 574 universities observed over 10 years (4,068 observations). Panel data analysis isolates individual effects for each university before estimating the coefficients of the explanatory variables, some of which vary over time. It also controls at the entity level for time-invariant omitted variables. Random effects method is chosen because the variation and differences across entities/universities, i.e., unique university characteristics are assumed to be random because they might have some influence on dependent variable. Fixed effects cannot be implemented because of the lack of variation in dummy independent variables. In the fixed effects model these variables are absorbed by the intercept. In addition, I performed a Sargan-Hansen test, which revealed that a RE estimator would have been a better choice than a FE.

The estimated panel regression equation is as follows:

$$Y_{it} = \alpha + \beta X_{it} + Z_i + u_{it} \text{ , where}$$

Y_{it}: overall university CPPS for university i in year t;
X_{it}: p-score for university i in year t, continuous;
Z_i: controls;
u_{it}: error term.

This formula allowed to answer the question of whether there a significant relationship between Russian university research productivity and the fact that its leader is a top scholar. To more fully answer this question, I interpret the results of panel regression analysis shown in Table 2 below.

Having controlled for the rector's sociodemographic characteristics, his/her experience, specialization, and term length, as well as university size, age, and participation in the 5-100 project, I found *no significant positive relationship between the variables of interest (column 1).* Basic models without controls (not reported here) demonstrate that there is a positive significant relationship between university productivity and the leader's prominence, but I have found that this relationship is *not* statistically significant when university size and the rector's specialization are controlled. This may be because these variables explain some variation in university productivity.

Table 2. Model 1. Panel regression results. Column 1: overall CPPS of the university 2010-2019 regressed on whether rector is a top scholar (measured by leader's p-score); Column 2: overall CPPS of the university 2010-2019 regressed on whether rector is a top scholar controlling for whether rector is an external candidate (dummy variable).

VARIABLES	*(1)*	*(2)*
p-score	7.926	7.988
	(5.228)	(5.304)
Sex	-46.90	-49.67*
	(28.73)	(29.98)
Term	4.310***	4.573***
	(1.445)	(1.501)
year_pub	1.201*	1.107
	(0.725)	(0.746)
N_authors	0.301***	0.301***
	(0.00603)	(0.00615)
project5100	160.6***	162.6***
	(56.82)	(57.98)
year_founded	0.293	0.235
	(0.257)	(0.267)
external_rector		-5.671
		(22.56)
Constant	-3,192**	-2,905*
	(1,544)	(1,593)
Observations	4,068	3,947
Number of universities	574	552
Controls for rector's specialization	YES	YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: Output for rector's specialization area is not included; model was tested on panel data using random effects.

Notably, participation in the 5-100 Project increases CPPS by more than 100 points, which can be evidence that research universities have a higher research productivity rate than the rest of the sample. The models control for the fact that university is a 5-100 project participant to get more reliable results. I've run a regression with an interaction between 5-100 Project universities and rector's p-score values which showed statistically significant and positive results (not reported here). Additionally, I've run a separate regression for 5-100 Project universities only meaning that there is a significant correlation between university research productivity and whether the university is a 5-100 Project participant. The Project created a competitive environment for university participants to score better in research in Russia and internationally. Average CPPS among 5-100 Project participants is 1473 in contrast to non-participants which equals 175. This means that research universities score 8 times higher in research than non-research. The results also demonstrate that term length is positively associated with overall CPPS. This may be because a longer term allows for the stability needed to achieve the results required for the university to be effective in research.

Column 2 in Table 2 presents panel regression results that examine the relationship between top leadership and overall university performance controlling for rector type. Internal rectors are those who have worked at the university prior to an appointment while external rectors are those who have not. It is of interest that the average CPPS for universities led by external rectors in the sample is 214.6 while for universities with internal rectors it equals 232.3. I found no observable relationship between overall university research performance and whether or not the rector is an external candidate. While external rector control is not statistically significant, the coefficients remain stable. These results do *not* support the posed hypotheses H1a and H1b. Firstly, there is no significant positive relationship between the academic success of a rector and the future research performance of the university in general. Secondly, a positive relationship between the academic success of a rector and the future research performance of the university in general for externally recruited leaders was *not* found.

Second model results: university research areas' productivity and top leadership

The second model seeks to examine whether the interaction between a rector's academic achievement and his/her specialization has an effect on that university research field's productivity. The model deals with various dependent variables, i.e., *CPPS for specific research fields* at the university (CPPS in mathematics, physics, etc.). These dependent variables were regressed on an interaction between *p-score* and whether the rector specializes in this particular area of research since I assumed that top leadership adds to research productivity. Importantly, the model requires that I consider all universities that have different departments and produce academic output for these fields, i.e., universities in the sample have multiple faculties, not only mathematics, physics, etc. Controls are the same as before, except for the *rector's specialization area* control, which was excluded as I use instead the rector's specialization dummy in the interaction. However, the *university type* dummy has been included instead to control for this.

The regression equation is as follows:

$$Y_{it} = \alpha + \beta_1 X1_{it} + \beta_2 X2_{it} + \beta_3 X1 * X2_{it} + \beta_4 Z + u_{it} + \varepsilon_{it} \text{ , where}$$

Y_{it}: CPPS for specific research fields, e.g., mathematics, for university i in year t;

$X1$: p-score, continuous;

$X2$: rector's specialization area, binary, for example, mathematics;

$X1*X2_{it}$: interaction between X1 p-score (continuous) and X2 (rector's specialization area, binary, e.g., mathematics)

Z: controls;

u_{it}: between entity error;

ε_{it}: within-entity error.

Table 3 shows the effects of the interaction between a rector's specialization and whether the rector is a top researcher for each research field. The results show that interactions between the two *is statistically significant for the research fields of physics, technical sciences, medicine, agricultural, and social sciences.* For instance, in the field of technical sciences, 7.7 is the change in CPPS if the rector is a technical sciences specialist and a top scholar in contrast to a vice versa situation. The interaction seems to bring about a greater change for the CPPS in medicine in comparison to other research areas. For every 1-unit increase in p-score, the CPPS in medicine should be expected to decrease by 21 points if the rector is not a specialist in the area, while otherwise it is expected to add an additional 321 points to the CPPS. On the other hand, I found that the interaction is *not significant for computer science, chemistry, earth sciences, humanities, and biology.* This may be related to the fact that some research fields lack data in general. Yet it can be confidently said that when rector is a specialist in the area of mathematics, computer, technical and social sciences there is a positive influence in the CPPS of this respective field. Moreover, I found that the longer a rector serves as rector is important for the CPPS in mathematics, physics, chemistry, earth sciences, and biology. All findings also suggest that larger universities tend to have higher CPPSs in all research fields.

I then run a three-way interaction to test whether there is an effect on the CPPS in research fields depending on whether rector is internal. Table 4 demonstrates that the CPPS in physics, agriculture, medicine, technical, and social sciences remains statistically related to an interaction between p-score and the rector's specialization controlling for whether the rector is an internal candidate. The statistically significant coefficients remain largely unaffected. These results partially support hypotheses H2a and H2b. Indeed, I find that for some research areas there is a positive relationship between the academic success of a rector and the university performance in the field in which the leader specializes. Likewise, the result stays significant when I add internal rector variable to an interaction.

Discussion

The results demonstrate that if a university leader is a successful academic, then s/he will positively influence university productivity only in specific research areas depending on his/her area of specialization. While the analysis did not support the hypotheses originating in NPM theory, I did find some evidence supporting the "political model" of leadership which states that politics is effective in enhancing research performance in particular research fields the rector is closest to. Paradoxically, a rector who does *not* prioritize overall university performance above department performance seems to be a more productive strategist in increasing productivity than the NPM strategic thinker who reallocates scarce resources within the university in a fair manner.

While the results shed some light on the nature of the relationship between university productivity and top leadership, we can also think about other possible explanations. For instance, universities might be considered as loosely coupled organizations since academic organizations embark on weak linkages between professionals in different areas rather than a tightly linked groups of professionals (Weick 1976). Universities embrace various disciplines so the content within the organization differs widely especially since these disciplines are weakly related to each other. In other words, universities differ in their internal organization compared to the typical organization as studied by NPM. This supports the finding that internal candidates can be better leaders (at least at some university types) since they possess incentives to benefit their respective subunits in comparison to external appointees. Hence, better connection to a discipline may be a reason why there is a relationship between top leadership and university productivity in certain areas.

In general, the results raise questions about the effectiveness of the NPM narrative, which has been influential in many countries that have sought to increase government efficiency. While NPM pretends to be the dominant policy in HE, it is also debated whether it is an effective one. The Russian case and several other empirical cases allow to conclude that universities—as specific organization type—require a special type of professional leadership (Goodall 2009; Engwall, Levay, and Lidman 1999). Academic excellence is the prerequisite for improving institutional performance in universities. Leaders with excellent academic records influence the academic performance in the research fields that are closest to. Nevertheless, there is a positive relationship between academic excellence in leadership and university performance.

References

Badillo-Vega, Rosalba, Georg Krücken, and Pedro Pineda. 2019. "Changing Analytical Levels and Methods of Leadership Research on University Presidents." *Studies in Higher Education*, 1–13.

Baldridge, J. Victor. 1971. "Models of University Governance: Bureaucratic, Collegial, and Political."

Bayers, Nancy K. 2005. “Using ISI Data in the Analysis of German National and Institutional Research Output.” *Scientometrics* 62 (1): 155–163.

Birnbaum, Robert. 1989. “Presidential Succession and Institutional Functioning in Higher Education.” *The Journal of Higher Education* 60 (2): 123–135.

Bornmann, Lutz, Rüdiger Mutz, Christoph Neuhaus, and Hans-Dieter Daniel. 2008. “Citation Counts for Research Evaluation: Standards of Good Practice for Analyzing Bibliometric Data and Presenting and Interpreting Results.” *Ethics in Science and Environmental Politics* 8 (1): 93–102.

Breakwell, Glynis M., and Michelle Y. Tytherleigh. 2008. “UK University Leaders at the Turn of the 21st Century: Changing Patterns in Their Socio-Demographic Characteristics.” *Higher Education* 56 (1): 109–127.

———. 2010. “University Leaders and University Performance in the United Kingdom: Is It ‘Who’Leads, or ‘Where’They Lead That Matters Most?” *Higher Education* 60 (5): 491–506.

Cannella Jr, Albert A., and Michael Lubatkin. 1993. “Succession as a Sociopolitical Process: Internal Impediments to Outsider Selection.” *Academy of Management Journal* 36 (4): 763–793.

Capano, Giliberto, Andrea Pritoni, and Giulia Vicentini. 2019. “Do Policy Instruments Matter? Governments’ Choice of Policy Mix and Higher Education Performance in Western Europe.” *Journal of Public Policy*, 1–27.

Carvalho, Teresa, and Rui Santiago. 2010. “Still Academics after All….” *Higher Education Policy* 23 (3): 397–411.

Clark, Burton R., and Ted IK Youn. 1976. “Academic Power in the United States: Comparative Historic and Structural Perspectives. Research Report No. 3.”

Cohen, Michael D., and James G. March. 1974. “Leadership and Ambiguity: The American College President.”

Decramer, Adelien, Carine Smolders, Alex Vanderstraeten, and Johan Christiaens. 2012. “The Impact of Institutional Pressures on Employee Performance Management Systems in Higher Education in the Low Countries.” *British Journal of Management* 23: S88–S103.

Deem, Rosemary, and Kevin J. Brehony. 2005. “Management as Ideology: The Case of ‘New Managerialism’in Higher Education.” *Oxford Review of Education* 31 (2): 217–235.

Diefenbach, Thomas. 2009. “New Public Management in Public Sector Organizations: The Dark Sides of Managerialistic ‘Enlightenment.’” *Public Administration* 87 (4): 892–909.

Dyachenko, Ekaterina, and Asya Mironenko. 2019. “Academic Leadership Through the Prism of Managerialism: The Relationship Between University Development and Rector’s Specialization.” *Educational Studies Moscow*, no. 1.

Engwall, Lars. 2014. “The Recruitment of University Top Leaders: Politics, Communities and Markets in Interaction.” *Scandinavian Journal of Management* 30 (3): 332–343.

Engwall, Lars, Charlotta Levay, and Rufus Lidman. 1999. “The Roles of University and College Rectors.” *Higher Education Management* 11: 75–94.

Etzioni, Amitai. 1959. “Authority Structure and Organizational Effectiveness.” *Administrative Science Quarterly*, 43–67.

Fligstein, Neil. 1987. “The Intraorganizational Power Struggle: Rise of Finance Personnel to Top Leadership in Large Corporations, 1919-1979.” *American Sociological Review*, 44–58.

Goodall, Amanda H. 2005. “Should Top Universities Be Led by Top Researchers and Are They? A Citations Analysis.”

———. 2009. “Highly Cited Leaders and the Performance of Research Universities.” *Research Policy* 38 (7): 1079–1092.

Goodall, Amanda H., John M. McDowell, and Larry D. Singell. 2017. “Do Economics Departments Improve after They Appoint a Top Scholar as Chairperson?” *Kyklos* 70 (4): 546–564.

Gumport, Paricia J. 2000. “Academic Restructuring: Organizational Change and Institutional Imperatives.” *Higher Education* 39 (1): 67–91.

Henkel, Mary. 2002. “Emerging Concepts of Academic Leadership and Their Implications for Intra-Institutional Roles and Relationships in Higher Education.” *European Journal of Education* 37 (1): 29–41.

Huang, Futao. 2017. “Who Leads China’s Leading Universities?” *Studies in Higher Education* 42 (1): 79–96.

Julius, Daniel J., J. Victor Baldridge, and Jeffrey Pfeffer. 1999. “A Memo from Machiavelli.” *The Journal of Higher Education* 70 (2): 113–133.

Kraatz, Matthew S., and James H. Moore. 2002. “Executive Migration and Institutional Change.” *Academy of Management Journal* 45 (1): 120–143.

Muller, Jerry Z. 2018. *The Tyranny of Metrics*. Princeton University Press.

Musselin, Christine. 2007. “Are Universities Specific Organisations.” *Towards a Multiversity*, 63–84.

Muzzin, Linda J., and George S. Tracz. 1981. “Characteristics and Careers of Canadian University Presidents.” *Higher Education* 10 (3): 335–351.

Parker, Lee D. 2013. “Contemporary University Strategising: The Financial Imperative.” *Financial Accountability & Management* 29 (1): 1–25.

Pfeffer, Jeffrey. 1981. *Power in Organizations*. Vol. 33. Pitman Marshfield, MA.

Pfeffer, Jeffrey, and William L. Moore. 1980. “Power in University Budgeting: A Replication and Extension.” *Administrative Science Quarterly*, 637–653.

Pfeffer, Jeffrey, and Gerald R. Salancik. 1974. “Organizational Decision Making as a Political Process: The Case of a University Budget.” *Administrative Science Quarterly*, 135–151.

Rowley, Daniel James, and Herbert Sherman. 2003. “The Special Challenges of Academic Leadership.” *Management Decision*.

Salancik, Gerald R., and Jeffrey Pfeffer. 1974. “The Bases and Use of Power in Organizational Decision Making: The Case of a University.” *Administrative Science Quarterly*, 453–473.

Sīle, Linda, Janne Pölönen, Gunnar Sivertsen, Raf Guns, Tim CE Engels, Pavel Arefiev, Marta Dušková, Lotte Faurbæk, András Holl, and Emanuel Kulczycki. 2018. “Comprehensiveness of National Bibliographic Databases for Social Sciences and Humanities: Findings from a European Survey.” *Research Evaluation* 27 (4): 310–322.

Singell Jr, Larry D., and Hui-Hsuan Tang. 2013. “Pomp and Circumstance: University Presidents and the Role of Human Capital in Determining Who Leads US Research Institutions.” *Economics of Education Review* 32: 219–233.

Sokolov, Mikhail, Sofia Lopatina, and Gennady Yakovlev. 2018. “From Partnerships to Bureaucracies: The Constitutional Evolution of Russian Universities.” *Educational Studies Moscow*, no. 3 (eng).

Staw, Barry M. 1980. “The Consequences of Turnover.” *Journal of Occupational Behaviour*, 253–273.

Taylor, Jim. 2011. “The Assessment of Research Quality in UK Universities: Peer Review or Metrics?” *British Journal of Management* 22 (2): 202–217.

Thornton, Patricia H., and William Ocasio. 1999. “Institutional Logics and the Historical Contingency of Power in Organizations: Executive Succession in the Higher Education Publishing Industry, 1958–1990.” *American Journal of Sociology* 105 (3): 801–843.

Tight, Malcolm. 2004. “Research into Higher Education: An a-Theoretical Community of Practice?” *Higher Education Research & Development* 23 (4): 395–411.

Weick, Karl E. 1976. “Educational Organizations as Loosely Coupled Systems.” *Administrative Science Quarterly*, 1–19.

Yielder, Jill, and Andrew Codling. 2004. “Management and Leadership in the Contemporary University.” *Journal of Higher Education Policy and Management* 26 (3): 315–328.

Zarate, Romualdo Lopez. 2007. “Four Trajectories of Rectors in Mexican Public Universities.” *Higher Education* 54 (6): 795–817.

Zhang, Yan, and Nandini Rajagopalan. 2004. “When the Known Devil Is Better than an Unknown God: An Empirical Study of the Antecedents and Consequences of Relay CEO Successions.” *Academy of Management Journal* 47 (4): 483–500.

Table 4. Model 2. Panel regression results. CPPS in specific research fields regressed on an interaction between rector's p-score (p-score), whether rector specializes in this specific area of research (cat_rector), and whether rector is an internal candidate (internal_rector).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
VARIABLES	mathematics	computer science	physics	chemistry	earth	biology	technical	medicine	agriculture	social	humanities
cat_rector	35.71	12.60***	3.707	39.38	8.953	-7.223	13.41	40.48*	1.078	29.68**	3.409
	(30.62)	(2.814)	(50.44)	(94.22)	(17.10)	(17.92)	(17.03)	(23.23)	(4.290)	(14.66)	(7.619)
p-score	10.92***	1.070***	7.640	-3.591	0.277	0.186	1.972	2.686	0.492	3.437	10.04***
	(2.831)	(0.366)	(9.697)	(8.148)	(1.479)	(4.166)	(4.284)	(3.287)	(1.273)	(13.29)	(3.022)
cat_rector#p-score	188.7	-19.67	2,257***	209.0	0.822	48.75	388.8***	1,696***	8.599***	76.78***	24.08
	(254.3)	(25.79)	(422.4)	(3,839)	(99.52)	(31.95)	(32.92)	(184.4)	(2.029)	(17.04)	(47.73)
internal_rector	1.894	0.0904	3.734	7.316	2.180	2.894	4.635	-2.113	1.447	3.370	1.612
	(2.848)	(0.299)	(8.204)	(7.837)	(1.673)	(3.824)	(5.412)	(4.808)	(1.300)	(11.86)	(3.045)
cat_rector#internal_rector	52.34	-13.62***	-45.46	-22.18	-17.88	.	23.58	-92.59***	3.044	-20.72	7.414
	(33.25)	(3.808)	(59.32)	(93.98)	(18.53)	(.)	(17.98)	(19.29)	(4.354)	(16.40)	(8.439)
internal_rector#p-score	-10.95***	-1.042***	-6.596	4.648	0.0754	0.201	-1.497	-2.457	-0.496	-2.549	-9.438***
	(2.870)	(0.372)	(9.851)	(8.264)	(1.567)	(4.228)	(4.342)	(3.327)	(1.284)	(13.36)	(3.065)
cat_rector#internal_rector#p-score	-478.1*	26.55	7,929***	-268.1	-1.514	.	383.0***	1,382***	7.377***	44.15**	-9.799
	(255.9)	(26.58)	(1,815)	(3,851)	(99.52)	(.)	(33.03)	(187.2)	(2.081)	(22.09)	(49.03)
sex	-4.261	-0.184	-20.34**	-12.82	-2.211	-4.819	-5.432	-4.180	-1.207	14.40	-0.807
	(3.532)	(0.372)	(10.04)	(9.728)	(2.082)	(4.764)	(6.428)	(5.797)	(1.562)	(10.38)	(3.553)
term	1.158***	0.0142	1.609***	1.360***	0.293***	0.789***	-0.563	-0.0189	0.0619	-0.114	0.0509
	(0.189)	(0.0199)	(0.535)	(0.519)	(0.111)	(0.255)	(0.343)	(0.310)	(0.0828)	(0.548)	(0.190)
year_pub	0.289***	0.00888	1.185***	0.642***	0.137***	0.267**	0.151	-0.323**	0.00811	0.374	0.150*
	(0.0897)	(0.00945)	(0.268)	(0.248)	(0.0528)	(0.127)	(0.162)	(0.148)	(0.0394)	(0.263)	(0.0903)
N_authors	0.0273***	0.00102***	0.0521***	0.0533***	0.0107***	0.0260***	0.0218***	0.00919***	0.00345***	0.0623***	0.0220***
	(0.000854)	(8.16e-05)	(0.00224)	(0.00215)	(0.000461)	(0.00105)	(0.00142)	(0.00130)	(0.000345)	(0.00228)	(0.000787)
project5100	-37.77***	3.695***	60.81***	-105.6***	-26.64***	-51.36***	96.09***	57.50***	-9.762***	-0.754	-13.82*
	(7.471)	(0.766)	(21.15)	(20.11)	(4.320)	(9.900)	(13.26)	(12.12)	(3.220)	(21.14)	(7.353)
year_founded	0.110***	0.00245	0.177*	0.170*	0.0119	0.0717	0.0556	-0.240***	-0.0125	0.263***	-0.0552
	(0.0350)	(0.00367)	(0.0993)	(0.0971)	(0.0206)	(0.0474)	(0.0636)	(0.0578)	(0.0154)	(0.102)	(0.0353)
Constant	-795.6***	-30.79	-2,706***	-1,622***	-297.6***	-677.7**	-388.0	1,130***	9.486	-1,289**	-188.0
	(194.5)	(20.46)	(575.3)	(539.0)	(114.5)	(272.2)	(351.5)	(320.1)	(85.40)	(573.6)	(196.0)
Observations	3,959	3,959	3,959	3,959	3,959	3,959	3,959	3,959	3,959	3,959	3,959
Number of universities	552	552	552	552	552	552	552	552	552	552	552
Control for university type (category_uni_encoded)	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1)

Notes: Output for university type dummy is not included; model was tested on panel data with random effects. Each research area corresponds to a column. Dots in 6th column indicate lack of observations.

Where international development and gender equality meet in science: a bibliometric analysis

Gita Ghiasi[1], Matthew Harsh[2], Tanja Tajmel[3] and Vincent Larivière[1]

[1] *gita.ghiasi.hafezi@umontreal.ca; vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, 3150 rue Jean-Brillant, Montréal, QC (Canada)

[2] *mharsh@calpoly.edu*
Interdisciplinary Studies in Liberal Arts, California Polytechnic State University, San Luis Obispo, CA 93407-0329 (USA)

[3] *tanja.tajmel@concordia.ca*
Centre for Engineering in Society, Concordia University, 1455 De Maisonneuve Blvd. W., Montréal, QC. H3G 1M8 (Canada)

Abstract

This study examines how international development and gender equality reinforce one another through the introduction and implementation of mechanisms to (1) promote women's engagement in scientific and technological advancements that hold potent promises for developing countries and the attainment of United Nation's (UN) sustainable development goals (SDGs), and to (2) orient more scientific activities toward promoting gender equality (SDG5). It addresses these two dimensions and answers three main questions: (1) to what extent scientific efforts are aligned with the achievement SDGs, (2) where SDG5 (Gender equality) stands in these scientific efforts, and finally (3) who is involved in the scientific advancements that contribute to the success of SDGs and SDG5, in terms of disciplines, countries, economies, and gender. The findings call for the need to support more research in relation to SDGs, particularly SDG5, which could potentially reinforce the engagement of more women in science from both developing and developed nations. The results of this study thus have great implications for both developed and developing nations and underpin the importance of cross-dimensional analysis, which is often overlooked in retention and inclusion policies of women in S&T.

Introduction

Gender inequality prevails in science. Women comprise fewer than 30% of scientific authorship, and their publications receive lower citation rates (Larivière et al. 2013). That is largely perpetuated by the lack of women in science, technology, engineering, and mathematics (STEM) fields. However, this explanation is only partial. Other factors are (but not limited to): under-recognition of women's contributions to science in the receipt of prestigious awards and grants (Lincoln et al. 2012), gender biases in perceptions of publication quality and collaboration interest (Knobloch-Westerwick et al. 2013), persistent gender biases in evaluations of research work (Murray et al. 2019). Yet, the reward system of science revolves around scientific productivity, which coupled with citation measures could reinforce biases in resource distribution and salary (Hilmer et al. 2015), hiring, appointment, tenure, and promotion (Holden et al. 2005), and potentially leave women disadvantaged and underprivileged. This not only has consequences for gender equality but also for research outcomes. That is, the dearth of women in science could also be reflected in the lack of attention to gender-related factors in research and scientific discoveries (Nielsen et al. 2017)—also known as gendered innovation (Schiebinger 2014), which could potentially circumscribe research agendas and priorities.

In response to the growing concern on the dearth of women in science, science and technology policies largely focus on either increase in participation of women or on affirmative action practices to remedy the attrition of women from the science pipeline. However, these efforts could be futile, unless coupled with systematic efforts to alleviate inequalities in other dimensions that are likely to hinder social inclusion and cohesion (Cozzens et al. 2007). One primary implication is to promote international development through the introduction and implementation of mechanisms to (1) promote women's engagement in scientific and technological advancements that hold potent promises for developing countries and the attainment of United Nation's (UN) sustainable development goals (SDGs), and to (2) orient more scientific activities toward promoting gender equality (SDG5).

This study seeks to address these two priority dimensions, and answers three main questions: (1) to what extent scientific efforts are aligned with the achievement SDGs, (2) where SDG5 (Gender equality) stands in these scientific efforts, and finally (3) who is involved in the scientific advancements that contribute to the success of SDGs and SDG5, in terms of disciplines, countries, economies (world bank classification), and gender. The results of this study help underpin the need for the incorporation of cross-dimensional analysis into gender-related science and technology policies.

Methods

Data is extracted from the Web of Science (WoS) from the years 2008-2018. SDG-related papers—i.e., papers that are in line with the SDGs—are identified using the list of WoS keyword queries developed by the IRDG Office at the University of Toronto (University of Toronto 2020, p. 42). This list entails a set of detailed queries designed and developed for each goal for the analysis of research papers. 388,334 papers are identified over the years 2008-2018 that are associated with the SDGs. Each paper is assigned to only one discipline based on National Science Foundation's (NSF) journal classification. Citation measures (both citations and journal impact factor) are normalized by field and year of publication. Countries of affiliations are further classified into high income, upper-middle-income, lower-middle-income and low-income economies based on World Bank (WB) country classifications by income level (2018-2019). Gender of authors is assigned in a similar approach as in Larivière et al. (2013) by using universal and country-specific existing name and gender databases. The proportion of the scientific output of authors of each gender is defined as a fractional count of papers, according to which each author is given a 1/x count of authorships. This study considers first and corresponding authors as a proxy for lead authors who make the main contributions to a scientific paper (Larivière et al. 2016; Mattsson et al. 2011). Accordingly, when the *first and the corresponding authors* of a paper are of the *same gender, or of the same country, or of the same economy (WB country classification)*, this study considers that the main contributions to that paper are made by that gender, that country, or that economy. Table 1 shows the number of SDG-related papers by gender, country of affiliation, and WB country categories of the lead authors.

Table 1- Number of SDG-related papers by gender, country, and WB country categories of the lead author(s)

Gender	*# of papers*	*Top countries*	*# of papers*	*WB category*	*# of papers*	*Discipline*	*# of papers*
Female	92009	*USA*	112835	*High-income*	252,926	*Arts and humanities (AH)*	886
		China	57104	*Upper middle-income*	88,635	*Natural sciences & engineering (NSE)*	218949
Male	148621	*UK*	36956	*Lower middle-income*	17,391	*Biomedical research*	92099
		Australia	25133	*Low income*	1,583	*Social Sciences*	76400

Results

Our findings reveal a continuous increase in the share of SDG-related papers (as of the total number of papers published) over the period 2008-2018 (Figure 1). Overall, SDG-related papers form 2.73% of total publications in years 2008-2018, which accounts for 1.5%, 2.8%, and 4.6% of total papers published in the fields of biomedical research, natural sciences and engineering (NSE), and social sciences, respectively. The share of SDG-related papers in 2008-2018 follows a linear trend line, which shows that the increase in this share has been steady over the years and that the introduction of SDGs in 2015 did not provide an upsurge in research associated with the SDGs.

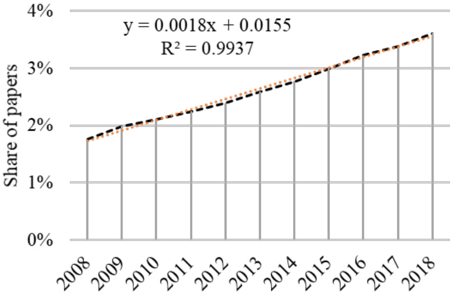


Figure 1. Share of SDG-related publications over years.

While examining SDG-related publications, our results show that clean water and sanitation (SDG6), climate action (SDG13), good health and well-being (SDG3), and life on land (SDG15) were among the top focus areas of SDG-related articles. Gender equality (SDG5) accounts for only 5% of SDG-related publications and falls into the 6th area of focus of SDG-related publications (Figure 2).

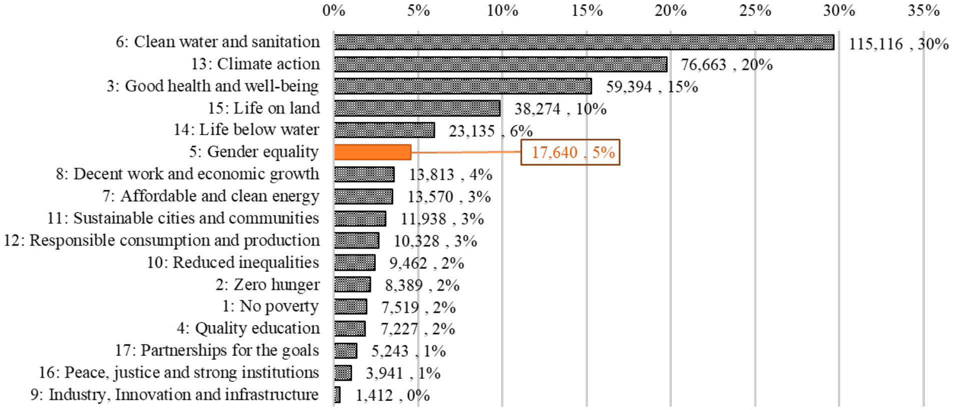


Figure 2. Number of and share of papers related to each SDG (to all SDG-related papers).

Our results highlight that among SDG-related papers, there has been a substantial increase in areas associated with climate change (SDG13) (increased from 13% in 2008 to 23% in 2018), and affordable and clean energy (SDG7) (increased from 1% in 2008 to 5% in 2018), and decrease in research focus areas associated with good health and well-being (19% in 2008 to 12% in 2018) (Figure 3). The share of papers associated with SDG5 (gender equality) stayed the same over the 11-years period. Our findings also show that the area of focus among SDG-related papers varies by country and economy (Figure 4a and 4b). The main focus of SDG-related papers published by high-income countries is on clean water and sanitation (SDG6, 24% of their SDG-related publications) and climate change (SDG13, 22% of their SDG-related publications), while for upper- and lower-middle-income countries is clean water and sanitation (SDG6, 43% of their SDG-related publications), and for low-income countries is good health and well-being (SDG3, 36% of their SDG-related publications). When looking into SDG5, our analysis reveals that the majority of SDG5-related papers are written by researchers from the US, the UK and Canada (Figure 4a) and that a higher proportion of SDG-related papers published by high income and low income countries is dedicated to SDG5 (Figure 4b).

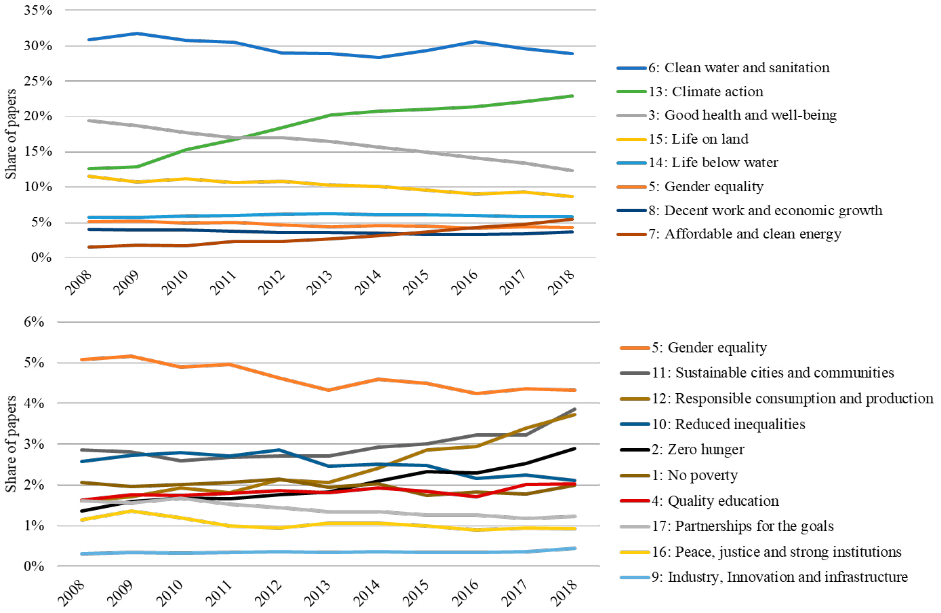


Figure 3. Share of papers related to each SDG (to all SDG-related papers) by publication year.

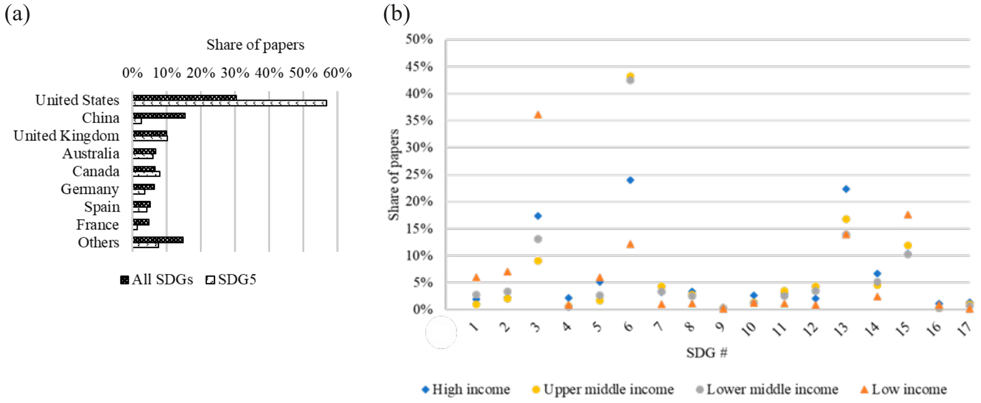


Figure 4. (a) Share of top prolific countries in papers related to all SDGs and SDG5 (left); and (b) Share of papers related to each SDG (to all SDG-related papers) by WB country classification (right).

Women account for 36.5% of total SDG-related authorship. The representation of women is more conspicuous in goals: gender equality (SDG5), and quality education (SDG4) (Figure 5). Moreover, women are highly represented in scientific activities related to SDG5 in all low, middle, and high-income countries (Figure 6a). However, despite women's higher representation, those papers that are led by women received lower citations and were published in lower impact factor journals (Figure 6b).

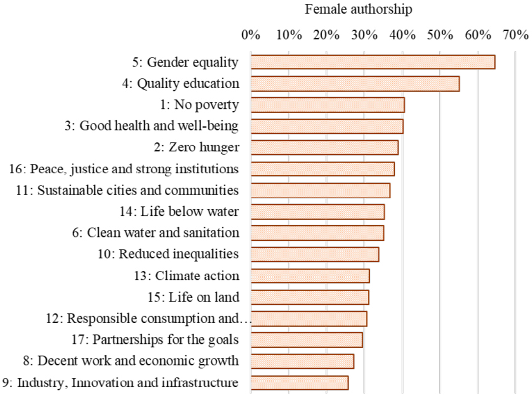


Figure 5. Share of female authorship by SDG.

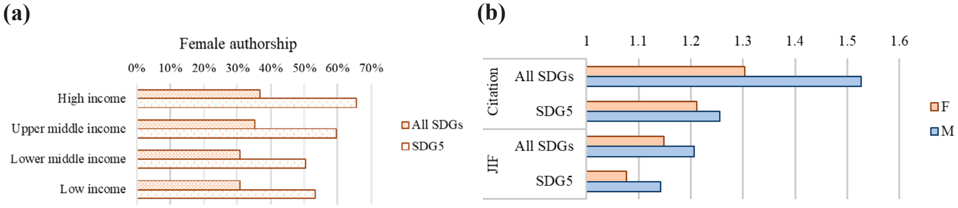


Figure 6. (a) share of female authorship in SDG-related papers and SDG5 by WB country classification; (b) relative citations received by papers led by women and men and the relative impact factor of the journals they are published in.

Conclusion

This study examines scientific publications that contribute to the attainment of SDGs and determines which goals are of primary importance to these scientific activities globally. It also identifies the relationship between lead authors' characteristics (gender, country, and economy of affiliation) with scientific efforts associated with each SDG. It also pays particular attention to SDG5 (gender equality) and maps to what extent these publications are led by women.

Our results show that the top focus areas of SDG-related research are different by the level of income of countries. High-income countries pay more attention to clean water and sanitation (SDG6) and climate change (SDG13), while upper-and-lower middle-income countries focus more on clean water and sanitation (SDG6), and low-income countries focus highly on good health and well-being (SDG3). The majority of SDG5-related papers are published by researchers from the US, the UK and Canada. China is an active player in SDG-related publications, but not in SDG5.Women represent 36% of total SDG-related authorship, which is higher than the representation of women in scientific authorship across all disciplines (i.e., 30% (Larivière et al. 2013). Women are highly represented in the scientific research related to

SDG5 (gender equality) regardless of their WB country classification. However, their publications are published in journals with lower impact factors and receive lower citation rates. Finally, our study shows that there has been no substantial increase in the share of papers related to SDG5 over eleven years (even after the proposition of SDGs in 2015).

These findings call for the need to support more research in relation to SDGs, particularly SDG5, which could positively reinforce the engagement of more women in science from both developing and developed nations. The results of this study thus have great implications for both developed and developing countries and underpin the importance of cross-dimensional analysis (looking at equity in terms of participation and outcomes of science), which is often overlooked in retention and inclusion policies of women in S&T. This study can advise on international development policies, including (1) encouraging women's participation in scientific research and promoting collaboration with women, and (2) providing opportunities for women to stay active in research and facilitating their ascent to the scientific eminence.

References

Cozzens, S., Kallerud, E., Ackers, L., Gill, B., Harper, J., Pereira, T. S. & Zarb-Adami, N. (2007). *Problems of Inequality in Science, Technology, and Innovation Policy*. James Martin Institute Working Paper 5. Oxford, UK.

Hilmer, M. J., Ransom, M. R. & Hilmer, C. E. (2015). Fame and the fortune of academic economists: How the market rewards influential research in economics. *Southern Economic Journal*, 82(2), 430–452. https://doi.org/10.1002/soej.12037

Holden, G., Rosenberg, G. & Barker, K. (2005). Bibliometrics: A potential decision making aid in hiring, reappointment, tenure and promotion decisions. *Social work in health care*, 41(3–4), 67–92.

Knobloch-Westerwick, S., Glynn, C. J. & Huge, M. (2013). The Matilda Effect in Science Communication: An Experiment on Gender Bias in Publication Quality Perceptions and Collaboration Interest. *Science Communication*, 35(5), 603–625. https://doi.org/10.1177/1075547012472684

Larivière, V., Desrochers, N., Macaluso, B., Mongeon, P., Paul-Hus, A. & Sugimoto, C. R. (2016). Contributorship and division of labor in knowledge production. *Social Studies of Science*, 46(3), 417–435. https://doi.org/10.1177/0306312716650046

Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, 504(7479), 211–213. https://doi.org/10.1038/504211a

Lincoln, A. E., Pincus, S., Koster, J. B. & Leboy, P. S. (2012). The Matilda Effect in science: Awards and prizes in the US, 1990s and 2000s. *Social Studies of Science*, 42(2), 307–320. https://doi.org/10.1177/0306312711435830

Mattsson, P., Sundberg, C. J. & Laget, P. (2011). Is correspondence reflected in the author position? A bibliometric study of the relation between corresponding author and byline position. *Scientometrics*, 87(1), 99–105. https://doi.org/10.1007/s11192-010-0310-9

Murray, D., Siler, K., Larivière, V., Chan, W. M., Collings, A. M., Raymond, J. & Sugimoto, C. R. (2019). Gender and international diversity improves equity in peer review. *BioRxiv*, 400515.

Nielsen, M. W., Andersen, J. P., Schiebinger, L. & Schneider, J. W. (2017). One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis. *Nature Human Behaviour*, 1(11), 791–796. https://doi.org/10.1038/s41562-017-0235-x

Schiebinger, L. (2014). Gendered innovations: harnessing the creative power of sex and gender analysis to discover new ideas and develop new technologies. *Triple Helix*, 1(1), 9. https://doi.org/10.1186/s40604-014-0009-7

University of Toronto. (2020). *Sustainable Development Goals at the University of Toronto* (p. 46). https://data.utoronto.ca/wp-content/uploads/2020/02/SUSTAINABLE-DEVELOPMENT-conceptual-report.pdf. Accessed 14 February 2021

Improving the precision of subject assignment for disparity measurement in studies of interdisciplinary research

Wolfgang Glänzel[1], Bart Thijs[2] and Ying Huang[3]

[1]*wolfgang.glanzel@kuleuven.be,* [2]*bart.thijs@kuleuven.be,* [3]*ying.huang@kuleuven.be*
KU Leuven, ECOOM & Dept MSI, Leuven Belgium

Abstract

Studies of interdisciplinarity research (IDR) poses severe challenges to bibliometricians. These challenges range from conceptualisation of IDR, over the definition of disciplines and the way of how research can be assigned to those, to finding the particular methods for quantifying and measuring the peculiarities of IDR. One of the key issues is the determination of granularity. This issue is twofold: The conceptional consideration should clarify the question the level at which IDR would be studied, namely as topic, subject or field interdisciplinarity, and this needs to be supported by quantitative results. The second key issue concerns the way of the assignment of the subjects that are integrated in the research, once the granularity level has been chosen. These two questions are tackled in the present paper, which is closely linked to further studies by the authors on the knowledge diffusion impact of IDR, on different implementations of similarity for the measurement of disparity and variety, and on the effect of similarity measurement approaches on indicators (see references below). The present study proposes a multiple-generation reference model and gives solutions for individual-document based subject assignment and the calculation of cognitive distances between disciplines needed for the determination of disparity measures.

Introduction

As science increasingly deals with boundary-spanning problems, important research ideas often transcend the scope of a single discipline or program. Thus, building sustainable bridges between two or more disciplines is valuable for pushing academic capability forward and for accelerating scientific discovery. Despite the growing attention interdisciplinary research (IDR) has received, there is a lack of consensus in the literature as to the definition of "interdisciplinary" (Huutoniemi et al., 2010). The definition of a "discipline" and discussions of the varieties of interdisciplinary, multidisciplinary, and transdisciplinary research have occupied much scholarly debate (NSF, 2004).

As a multi-faceted concept, IDR can mean different things to different people. One of the most broadly-accepted definitions of IDR is set forth in a National Academies' report (COSEPUP, 2004):

> *"a mode of research by teams or individuals that integrates information, data, techniques, tools, perspectives, concepts and/or theories from two or more disciplines or bodies of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or area of research practice."*

Derived from the above definitions, knowledge integration is the essence of IDR. According to Porter et al., (2008), the knowledge to be integrated can be of various forms: ideas (such as concepts and theories), methods (techniques and tools), or data from various fields of knowledge. The research on interdisciplinarity can be approached from several different perspectives. From a conceptual viewpoint, there are two main perspectives of studying interdisciplinarity, namely the cognitive perspective on the basis of information flows across disciplines and the organisational approach on the basis of co-authors' professional skills and/or affiliation. The cognitive approach can be implemented by using two different methods that can actually be combined into a hybrid methodology: Since abstract and citation databases provide the necessary link and textual information on citation flows and document topics, respectively, on the large scale even providing the basis of benchmarking, the cognitive approach can be considered the first option. By contrast, useful large-scale information on skills or affiliation needed for the organisational approach is only covered partially (Zhang et al., 2018). On the other hand, subject classification schemes provided with the databases or based on journal assignment and fail in the context of interdisciplinarity whenever journals have a general or even multidisciplinary scope. In order to gain reliable information on the subject of individual documents indexed in the database, a document-based improvement of the subject assignment is needed. The question arises of whether such an assignment with appropriate granularity is feasible on the large scale. We propose a parameterized rule-based model, implement different versions and study the effect of altering parameter settings. The assignment of a publication is based on the most dominant subfield(s) in the aggregated set of cited references where not only the references in the publication are taken into consideration but also those included in the cited publications.

Conceptual considerations and research design

The two central concepts related to IDR are “diversity” and “coherence” (cf. Rafols and Meyer, 2010). While Stirling (1994) distinguished has three components of diversity *variety*, *balance*, and *disparity*, Rafols (2014) proposed to subdivide the notion of coherence into the three aspects *density*, *intensity* and *disparity*. In either component, but most notably in the context of variety and disparity, the correctness and the granularity of topic identification is crucial.

The first aim of the present study is therefore twofold, on the one hand, we attempt to choose a granularity level at which the identification of IDR has still a nuts-and-bolts use and interpretation (biomedical engineering or urbanism certainly represents a different level of IDR than stochastic geometry), secondly, we try to extend existing journal-based subject classification schemes towards the individual document-level assignment. In order to achieve this objective, we will first summarise the effect of granularity on subject disparity on the basis of literature cited by articles indexed in the 2018 volume of the Web of Science Core Collection. We have applied three scientometric link-based methods to measure similarity, bibliographic coupling, co-citation and cross-citation links.

In this context we note that in the present paper we just summarise the major outcomes of these two objectives, while we have devoted three separate studies to the knowledge diffusion impact of IDR (Huang et al., 2021a), the effect of the choice of the particular scientometric method (coupling, co- and cross-citation) on the definition of the (dis-)similarity measures (Thijs et al., 2021) and the analysis of the granularity level in the context of IDR (Huang et al., 2021b). We decided to publish these two important aspects in separate papers as the exhausting investigation would require more space than available in this paper. The main focus of the present study is actually laid on the assignment of individual documents to disciplines, as the journal-based assignment used by the database providers proved too unspecific and thus, at least in the case of more general and multidisciplinary journals, impractical for our purpose. Nonetheless, the choices of both granularity level and link-analysis are indispensable steps in quantification and measurement of IDR and need to be decided upon before (interlinked) document are properly assigned to subjects. We decided therefore, not to remove these steps but to report them giving additional examples to those already given in the two other studies by the authors.

The following sections describe the experiment and the major outcomes.

Data sources and data processing

All documents indexed as articles in the three journal editions (SCIE, SSCI, A&HCI) of the 2018 volume of the Web of Science (WoS) Core Collection have been extracted from the database. For evolutionary and robustness analyses the previous volumes covering the period

1999–2018 have been inclued. For the analysis of cited references, all items that are indexed in *any* WoS database are taken into consideration. As the cognitive subject scheme underlying this study, the adjusted Leuven-Budapest classification with 16 major and 74 sub-fields (disciplines) building upon the WoS Subject Categories (Glänzel et al., 2016) has been used. Figure 1 shows the subject structure of the Leuven-Budapest scheme at the sub-field level.

Methods and results

Determining the granularity level

The basic methodological idea was to conduct the research in several consecutive steps. The first one concerned the granularity proceeding from the assumption that the discriminative power of the level would not dramatically change when we adjust the journal-based assignment by an article-based version using the same hierarchic structure. We applied three scientometric methods based on bibliographic coupling (BC), co-citation (CO) and cross-citation (CC). Again, a systematic investigation of the appropriate methodology for link-based similarity measures has been conducted by Thijs et al. (2021). We here give an example to quantitatively underpin the choice of the granularity level, which, in turn, is necessary for the implementation of individual subject assignment. This, again, shows the interweaving of the tree aspects.

Since co-citations need an appropriate citation window, we have used the five-year window 2010–2014 in this case. Since we are proceeding from a vector-space model, a cosine measure was applied to determine the similarity of subjects. The cosine similarity is defined as the cosine of the angle between two vectors representing the respective documents, i.e., ratio of their scalar product and the product of their lengths. In the case of a Boolean vector space as, for instance, in the case of BC and CO, this reduces to Salton's measure, i.e., the ratio of the number of jointly cited/citing items and the geometric mean of the number of all cited/citing items. To cross-citations, we have used formula used by Zhang et al. (2010) in Eq. (1) and later applied in the context of IDR as well (cf. Zhang et al., 2016). Dissimilarities can readily be derived from the cosine measure by elementary arithmetic manipulation.

This step is required to select both the granularity level and the methodological basis for measuring the disparity aspect of subjects in possible IDR indicators. The correlation across granularity levels is strong (≥ 0.92 for major fields and disciplines and ≥ 0.82 for the subject categories and disciplines) but the slopes reveal systematic trends. These are also reflected by the means. Table 1 gives the *mean* and *minimum mean* similarity with other subjects according to the three approaches. BC provides in general stronger similarity than CC and CO at all three levels. Although the minima for the individual scientometric link-analyses are taken by different subjects, the deviation is not large because the values of G (Geo & space sciences) and L (Social sciences II), K6 (literature) and H2 (pure mathematics), and UT (poetry) and EO (classics), respectively are of similar order (EO = 0.018 in BC, H2 = 0.051 in CC and G = 0.216 in CO, which do, however, not appear in Table 1).

Table 1. Empirical similarity statistics at three granularity levels

	BC		**CC**		**CO**	
	Mean	*Minimum*	*Mean*	*Minimum*	*Mean*	*Minimum*
Major Field	0.364	0.248 (G)	0.326	0.180 (G)	0.314	0.208 (L)
Discipline	0.228	0.064 (H2)	0.180	0.038 (K6)	0.163	0.051 (H2)
Subject Category	0.139	0.007 (UT)	0.090	0.015 (EO)	0.071	0.006 (EO)

Data sourced from Clarivate Analytics Web of Science Core Collection

This confirms previous findings by Glänzel et al. (2009) concerning subject-normalised citation indicators according to which major fields proved too coarse and the lowest level, (subject categories) provide a fine-grained but very fuzzy subject assignment. Subfields could therefore serve as the favoured reference level for disciplines. This gives also some quantitative evidence to support the decision of not to go for kind of "topic interdisciplinarity" as sketched in the preliminary and more conceptual considerations above.

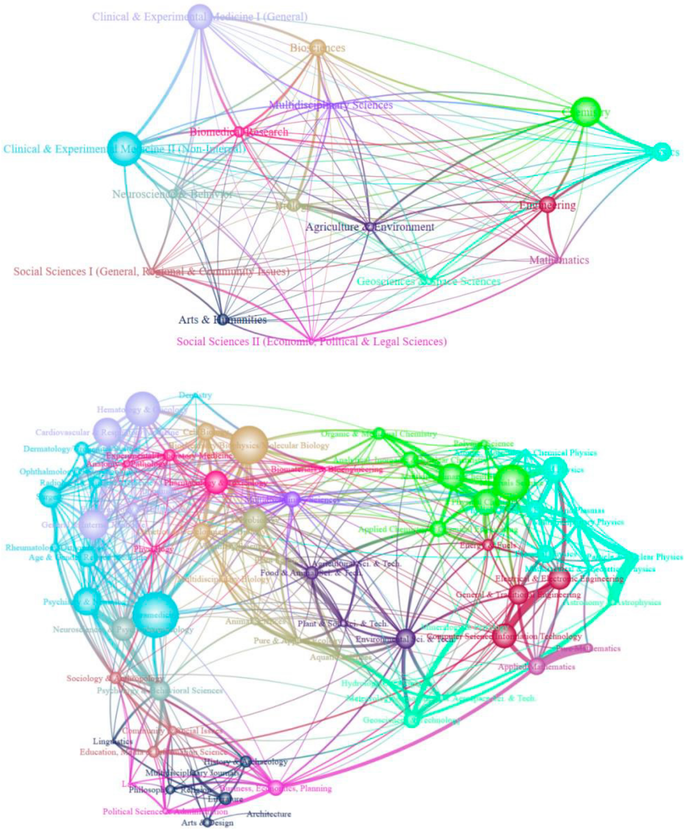

Clinical & Experimental Medicine I (General)
Biosciences
Multidisciplinary Sciences
Chemistry
Biomedical Research
Clinical & Experimental Medicine II (Non-Internal)
Physics
Neuroscience & Behavior
Biology
Engineering
Agriculture & Environment
Mathematics
Social Sciences I (General, Regional & Community Issues)
Geosciences & Space Sciences
Arts & Humanities
Social Sciences II (Economic, Political & Legal Sciences)
Dentistry
Hematology & Oncology
Cardiovascular & Respiratory Medicine
Cell Biology
Biochemistry/Biophysics/Molecular Biology
Organic & Medicinal Chemistry
Polymer Science
Chemical Physics
Dermatology/Urogenital System
Experimental/Laboratory Medicine
Anatomy & Pathology
Biomaterials & Bioengineering
Ophthalmology
Radiology
Surgery
Pharmacology & Toxicology
Multidisciplinary Sciences
Physical Chemistry
Plasmas
Physics
General & Internal Medicine
Applied Chemistry & Chemical Engineering
Energy & Fuels
Physiology
Rheumatology/Orthopedics
Age & Gender Related Medicine
Agricultural Sci. & Tech.
Food & Animal Sci. & Tech.
Particle & Nuclear Physics
Multidisciplinary Biology
Mathematical & Theoretical Physics
Electrical & Electronic Engineering
General & Traditional Engineering
Psychiatry & Neurology
Biomedicine
Astronomy & Astrophysics
Neuroscience & Psychopharmacology
Animal Sciences
Plant & Soil Sci. & Tech.
Mineralogy & Petrology
Pure & Applied Ecology
Environmental Sci. & Tech.
Computer Science/Information Technology
Pure Mathematics
Aquatic Sciences
Applied Mathematics
Sociology & Anthropology
Psychology & Behavioral Sciences
Hydrology
Meteorology
Aerospace Sci. & Tech.
Geosciences & Technology
Community & Social Issues
Linguistics
Education, Media & Information Science
History & Archaeology
Multidisciplinary Journals
Law
Philosophy & Religion
Business, Economics, Planning
Literature
Political Science & Administration
Architecture
Arts & Design

Figure 1. The cross-citation based structure of the Leuven-Budapest scheme (1999–2018) based on major fields (top) and subfields (bottom)

Data sourced from Clarivate Analytics Web of Science Core Collection

Figure 1 shows the disciplinary structure of the WoS at the broad field level (15 major fields) and subfield level (74 disciplines). The detailed scheme for these hierarchic levels is given in the Appendix.

Individual document based subject assignment

The next step towards creating the groundwork of variety and disparity measurement concerns the individual subject assignment of articles and their cited references. Variety (and balance) is based on the number and distribution of disciplines the knowledge of which is integrated in published research results, whereas disparity takes also their dissimilarity into account. While the first aspect can be studied by analysing, e.g., the disciplines to which the cited references belong, the later aspect requires the knowledge of the disciplinary structure of the complete database (cf. Figure 1). Because of the strong correlation and robustness of all three methods, any of those are suitable for the creation of a 'global' (dis-)similarity matrix. We decided to choose BC, since this does not require any particular citation window and most documents have sufficiently long reference lists. At this point, we have to make a distinction between the reference items used for BC, the particular topics of which are not relevant for the link analysis, and those used to improve subject assignment and to detect topics of knowledge integration in IDR. This forms a straight continuation and update of the idea proposed by Glänzel et al., (1999) and Glänzel & Schubert (2003) in connections with the creation of cognitive but bibliometrics-aided subject classification scheme. The proposed iterative algorithm is suited to assign documents to subjects at different levels of granularity, of course, with different precision. In this context we mention that Milojevič (2020) has recently developed a citation-based algorithm for to reclassify Web of Science articles at two different aggregation levels, to subject categories and broad disciplines.

Since many cited documents are published in general journals such as *Physical Review Letters* in physics, *JACS* or *Angewandte Chemie* in chemistry or even multidisciplinary journals like *Science*, *Nature*, *PNAS US* or *PLoS One* with no specific subject profile, we have to detect the topic of those items by analysing their own references. Therefore, we introduce the multiple-generation reference model to classify individual scientific publications. In this step, we adopt a full-counting method for the assignment classification system and we track the relationship between the original publication and the cited references' sub-fields (1st generation), but also the cited reference's sub-fields of those cited references (2nd generation) and so on (3rd generation).

The classification model we propose is a parameterized rule-based model. This approach allows us to implement different versions and study the effect of altering parameter settings. In short, a publication is assigned to the most dominant subfield(s) in the aggregated set of cited references. Cited subfields are ranked based on the normalized share they take in the total set across multiple generations. The selection of the most dominant subfields can be based on a particular threshold or a combination of rules or judgements.

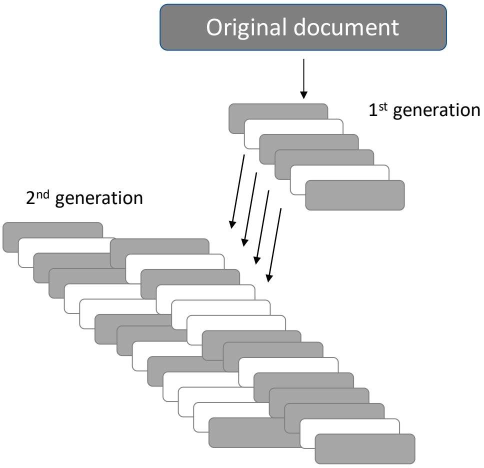


Figure 2. The multi-generation reference model supporting individual document subject-assignment based on two generations with 'active' (dark) and 'non-source' items (light).

Figure 2 illustrates the 2-generation approach. The grey-shaded references symbolise the "active" references, i.e. those that are indexed in the database, the white one stand for non-source references the assignment of which is often unclear and which are therefore ignored. Of course, each additional generation can increase the number of disciplines contributing to the integration of knowledge but could weaken their direct influences and thus increases fuzziness. Table 2 shows the distribution within the major fields of the similarity between the discipline profiles of the first two generations of cited references, where the share of the records within a certain similarity range in all records belonging to the corresponding field is calculated. According to the results, in some fields there is less integration of knowledge from other disciplines over reference generations than in other fields. The corresponding field codes can be found in the Appendix. The more skewed the distribution, the lower is the fuzziness and vice versa. Therefore, lower weights can be given to 'indirect' references, which are represented by higher generations and the precision of which in terms of the assignment of the original document decreases with the order of the generation.

Table 2. The distribution of discipline similarity between 1st and 2nd generation references by major fields, with a colour gradient from red (strong similarity) over white to blue (weak)

Field	[.95,1]	[.9, .95)	[.85, .9)	[.8, .85)	[.75, .8)	[.7, .75)	[.65, .7)	[.6, .65)	.55, 1.6)	[.5, .55)	[.0, .5)
A	0.456	0.268	0.120	0.062	0.034	0.020	0.012	0.008	0.006	0.004	0.009
B	0.484	0.282	0.119	0.054	0.026	0.014	0.008	0.004	0.003	0.002	0.003
C	0.510	0.249	0.109	0.054	0.029	0.017	0.010	0.007	0.004	0.003	0.007
E	0.578	0.199	0.090	0.047	0.027	0.019	0.011	0.008	0.006	0.004	0.011
G	0.729	0.139	0.055	0.028	0.015	0.010	0.007	0.005	0.004	0.003	0.007
H	0.640	0.164	0.076	0.041	0.023	0.020	0.010	0.007	0.006	0.004	0.010
I	0.467	0.280	0.124	0.058	0.029	0.016	0.009	0.006	0.004	0.002	0.005
K	0.368	0.160	0.108	0.077	0.050	0.061	0.032	0.026	0.029	0.023	0.065
L	0.692	0.144	0.064	0.033	0.019	0.015	0.009	0.006	0.005	0.004	0.010
M	0.501	0.241	0.113	0.058	0.032	0.019	0.012	0.008	0.005	0.004	0.008
N	0.675	0.195	0.067	0.029	0.014	0.008	0.005	0.003	0.002	0.001	0.003
P	0.535	0.228	0.101	0.052	0.029	0.018	0.011	0.008	0.005	0.004	0.009
R	0.339	0.295	0.162	0.086	0.047	0.027	0.016	0.010	0.006	0.004	0.008
Y	0.444	0.230	0.117	0.067	0.040	0.030	0.019	0.014	0.011	0.008	0.022
Z	0.484	0.262	0.113	0.056	0.031	0.019	0.011	0.007	0.005	0.004	0.008

Data sourced from Clarivate Analytics Web of Science Core Collection

The general formula of normalizing the share of a research field or discipline in our multiple-generation model is as follows:

$$\mathrm{NFS}^{(n)}{}_i = w_1\,\mathrm{FS}^{(1)}{}_i + w_2\,\mathrm{FS}^{(2)}{}_i + \ldots + w_{\mathrm{n}}\,\mathrm{FS}^{(n)}{}_i,$$

with

- n is the number of cited reference generations considered.
- $\mathrm{NFS}^{(n)}{}_i$ denotes the normalized share of category i aggregated over n generation.
- $\mathrm{FS}^{(k)}{}_i$ denotes the share of cited category i in the total number of cited categories at generation k.

- w_k refers to the normalizing factor or weight attributed to generation k. it takes a value between 0 and 1.
- $w_1 + w_2 + \ldots + w_n = 1$.

Based on the values that are assigned to the normalizing factors or weights special cases can be identified:

- If w_k=1 for any value of k where $1 \leqslant k \leqslant n$, only the k^{th} generation is taken into consideration.
- If w_k = 1/n for any value of k where $1 \leqslant k \leqslant n$, every generation is taken into consideration and treated equally.
- If $w_k = w_1/k$, for any value of k where $1 \leqslant k \leqslant n$, every generation is taken into consideration and the k^{th} generation has the $1/k$ times of share compared to the 1^{st} generation.
- If $w_k = w_1^k$, for any value of k where $1 \leqslant k \leqslant n$, each generation is taken into consideration and the factor for each additional generation is multiplied by w_1.

Next, the cited categories or fields are ranked in descending order based on their normalized shares. Finally, the highest ranked category or categories are attributed to the publication after the application of judgement rules. The normalized share for the highest ranked subfield is given by $\text{NFS}^{(n)}_{r=1}$

In this study, we consider only the first two generations and report on three weight-allocation schemes (or models) as shown in Table 3. References to *multi-disciplinary journals* in the last considered generation were ignored as being unspecific. The weights and their ratios have been determined on an empirical basis and simple arithmetic, where, for instance, M3 takes the potentiating number of references into account.

Table 3. The weight-allocation model for normalising the share of research fields

Wight model	Formula	w_1	w_2
M1	$w_1 = 1$	1.000	0.000
M2	$w_2 = 1$	0.000	1.000
M3	$w_2 = w_1^2$	0.618	0.382

For the assignment of the sub-fields to individual papers, we have implemented these selection rules:

- The individual assignment to a paper is limited to three sub-fields (i.e., disciplines) according to their frequency ranks i = 1, 2, 3.
- A publication can be uniquely assigned to discipline a only if none of the higher ranked subfields has a normalized share which is at most 0.67 times the subsequent one.
- Assignment of additional disciplines is done, if the above share is larger than 0.67 following the same algorithm.
- The procedure is to be stopped after at most three fields have been assigned. Otherwise, assignment is stopped by the procedure whenever $\text{NFS}^{(2)}_{r=i+1}/\text{NFS}^{(2)}_{r=i} \leqslant 2/3$ for any $i \leq 3$.
- Unassigned documents can still be assigned manually, but are very likely to be truly interdisciplinary themselves. These cased proved to be rather rare.

Table 4 provides a concise formalised view of the complete procedure of subfield assignment to individual papers.

We add a sample of the identified papers published in Nature to illustrate the procedure of assigning the field, shown in Table 5. Note, that the weight type here is M1, i.e., only considering the first-generation reference. As our purpose is to allocate up to three fields to the individual scientific publication, the fields labeled "Multidisciplinary Sciences (X0)" should be removed.

Table 4. Overview of the complete procedure of subfield assignment to individual papers

1st round judgement	2nd round judgement	3rd round judgement	Assignment
$NFS^{(2)}_{r=2}/NFS^{(2)}_{r=1} \leqslant 2/3$			Field 1
$2/3 < NFS^{(2)}_{r=2}/NFS^{(2)}_{r=1} \leqslant 1$	$NFS^{(2)}_{r=3}/NFS^{(2)}_{r=2} \leqslant 2/3$		Field 1 Field 2
$2/3 < NFS^{(2)}_{r=2}/NFS^{(2)}_{r=1} \leqslant 1$	$2/3 < NFS^{(2)}_{r=3}/NFS^{(2)}_{r=2} \leqslant 1$	$NFS^{(2)}_{r=4}/NFS^{(2)}_{r=3} \leqslant 2/3$	Field 1 Field 2 Field 3
$2/3 < NFS^{(2)}_{r=2}/NFS^{(2)}_{r=1} \leqslant 1$	$2/3 < NFS^{(2)}_{r=3}/NFS^{(2)}_{r=2} \leqslant 1$	$2/3 < NFS^{(2)}_{r=4}/NFS^{(2)}_{r=3} \leqslant 1$	Unassigned

Table 5. The sample to illustrate the procedure of assigning the field

ISI	Total Refs.	WoS Refs.	Field 1	Field 2	Field 3	Field 4	Assigned Field(s)
000419769300025	71	59	G2 (0.389)	X0 (0.254)	G4 (0.152)	G3 (0.085)	G2
000419769300035	37	35	Z3 (0.314)	B1 (0.229)	R3 (0.086)	B2 (0.057)	Z3; B1
000419769300037	59	58	X0 (0.268)	B2 (0.232)	B1 (0.214)	B3 (0.107)	X0; B2; B1
000419769300030	45	43	C6 (0.246)	C4 (0.174)	P1 (0.159)	X0 (0.145)	Unassigned
000419769300031	43	35	X0 (0.192)	P6 (0.154)	C4 (0.154)	C6 (0.154)	Unassigned

Data sourced from Clarivate Analytics Web of Science Core Collection

In this paper, we propose three ways to deal with the "Multidisciplinary Sciences (X0)" during the procedure. Samples are shown in Table 6.

Table 6. Comparative results of the procedure for individual subject assignment using different methods to resolve X0 assignment

Method	**Weight model**	**Rank1**		**Rank2**		**Rank3**		**Rank4**		**Assigned Field(s)**
		Share	Field	Share	Field	Share	Field	Share	Field	
Original	M1	0.268	X0	0.232	B2	0.214	B1	0.107	B3	X0; B2; B1
	M2	0.238	B2	0.219	X0	0.211	B1	0.107	I4	B2; X0; B1
	M3	0.247	B2	0.206	B1	0.139	X0	0.135	I4	B2; B1; X0
Remove X0 after original result	M1	-	-	0.232	B2	0.214	B1	0.107	B3	B2; B1
	M2	0.238	B2	-	-	0.211	B1	0.107	I4	B2; B1
	M3	0.247	B2	0.206	B1	-	-	0.135	I4	B2; B1
Remove X0 before calculating shares	M1	0.317	B2	0.293	B1	0.146	B3	0.122	I4	B2; B1
	M2	0.306	B2	0.272	B1	0.135	I4	0.125	B3	B2; B1
	M3	0.287	B2	0.239	B1	0.156	I4	0.09	B3	B2; B1
Remove X0 after calculating shares	M1	0.232	B2	0.214	B1	0.107	B3	0.089	I4	B2; B1
	M2	0.238	B2	0.211	B1	0.107	I4	0.096	B3	B2; B1
	M3	0.247	B2	0.206	B1	0.135	I4	0.077	B3	B2; B1

Data sourced from Clarivate Analytics Web of Science Core Collection

Table 7 gives an impression on the effect of the chosen formula and weight scheme on the share of possible individual assignment for four selected journals. Using the M2 and M3 alone usually resulted in lower shares, for instance, of about 0.87 for the journals *Nature*, *Science* and *PNAS*, only the combination of M1, M2 and M3 resulted in significant improvement of the share. The same applied to the general journals like JACS in chemistry but even in the more specific journals with already high w_1 scores, the (M1 + M2 + M3) combination resulted in further improvement (cf. JASIST in information and library science).

Table 7. Overview of the complete procedure of subfield assignment to individual papers

	Weight model	Original	Remove X0 from original result	Remove X0 before calculating shares	Remove X0 after calculating shares
Nature (X0) (8259)	M1	0.834	0.781	0.906	0.906
	M2	0.778	0.755	0.874	0.872
	M3	0.784	0.781	0.873	0.873
	(M1+M2)	0.861	0.825	0.927	0.927
	(M1+M2+M3)	0.913	0.890	0.956	0.956
JASIST (E1,Y1) (1788)	M1	0.967	0.961	0.971	0.971
	M2	0.938	0.933	0.942	0.942
	M3	0.915	0.911	0.924	0.924
	(M1+M2)	0.975	0.970	0.977	0.977
	(M1+M2+M3)	0.987	0.984	0.988	0.988
BOI (B0,B1,E1, H1,Z3) (7296)	M1	0.779	0.768	0.805	0.805
	M2	0.740	0.735	0.782	0.780
	M3	0.787	0.784	0.867	0.867
	(M1+M2)	0.845	0.837	0.870	0.868
	(M1+M2+M3)	0.927	0.923	0.952	0.952
JACS (C0) (28141)	M1	0.889	0.885	0.902	0.902
	M2	0.861	0.860	0.875	0.875
	M3	0.839	0.839	0.848	0.848
	(M1+M2)	0.912	0.909	0.920	0.920
	(M1+M2+M3)	0.940	0.939	0.945	0.945

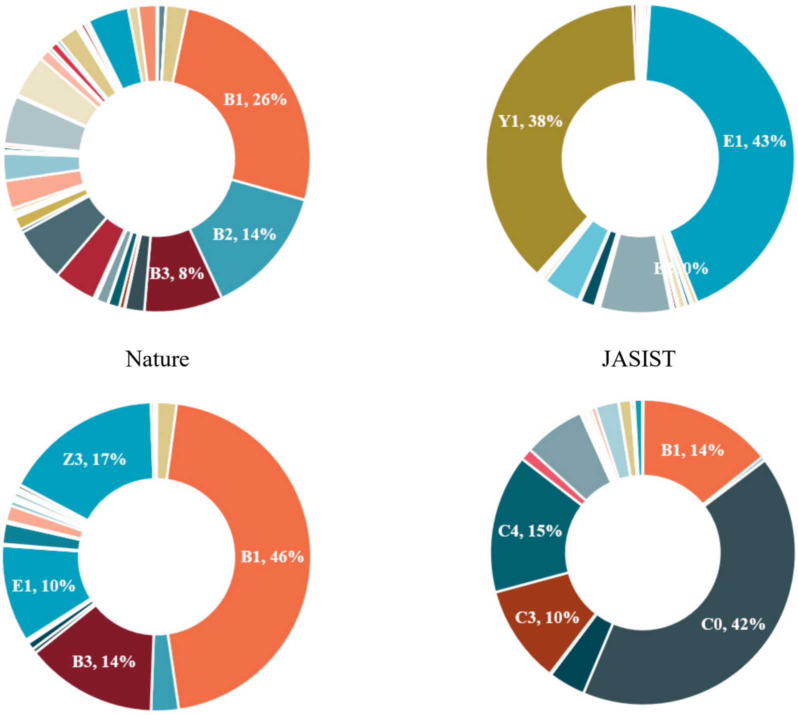
Data sourced from Clarivate Analytics Web of Science Core Collection

Our approach yields a similar precision as that proposed by Milojevič (2020): In fact, about 5% of articles published in general and multi-disciplinary journals could not be assigned individually. These papers proved highly interdisciplinary, mostly with no dominant discipline in the cited information sources so that the procedure, which was originally designed to be used in the context of IDR studies, helps directly identify interdisciplinary research quasi as by-product. Below we list four documents published in PLoS ONE and Nature just as typical examples of such research. Neither the cited references nor titles/abstracts and the authors' affiliations point to one specific discipline but immediately reveal the interdisciplinary nature of the underlying research:

- WOS:000383255900083 – Abundant topological outliers in social media data and their effect on spatial analysis (PLOS ONE)
- WOS:000391217400045 – Narrative style influences citation frequency in climate change science (PLOS ONE)
- WOS:000268938300034 – Dense packings of the Platonic and Archimedean solids (Nature)

- WOS:000376883000004 – The Kinome of pacific oyster Crassostrea Gigas, its expression during development and in response to environmental factors (PLOS ONE)

Figure 6 finally shows the subject profiles after individual subject assignment on the basis of the described iterative process of reference analysis. Above all, the three big multi-disciplinary journals reflect a large spectrum of (unequally distributed) disciplines. With the application of the models and methods described above, we have all prerequisites for the creation of the tools to measure the main aspects of IDR, most notably variety and disparity of integrated knowledge. However, this will be part of a separate study.



Nature

JASIST

Bioinformatics (BOI)

Journal of the American Chemical Society (JACS)

Figure 6. Visualisation of the subject profiles of several analysed multi- and interdisciplinary journals using 74 disciplines in the sciences, social sciences and humanities
Data sourced from Clarivate Analytics Web of Science Core Collection

Summary of main findings and conclusions

The experiment has provided three major results, firstly, bibliographic coupling, cross-citation and co-citations provided similar results in the analysis of subject (dis-)similarity of the WoS documents space, where BC proved a suitable and robust methodological basis at all granularity levels. Secondly, the granularity sub-field level with 74 disciplines in the sciences, socials sciences and humanities proved most suited for the large-scale analysis of IDR of individual documents and provides quantitative support to the conceptual consideration on what level the integration of knowledge may be studied. Thirdly and finally, the iterative process of weighted multi-generation reference analysis resulted in the applicability of the journals-based ECOOM classification scheme to individual-document assignment for ≥ 95% of documents. Recalling the objective of this study, namely improving the precision of disparity measurement in studies of interdisciplinary research, the achieved level of about 95% is sufficient: On the one hand, the method helps build reliable (dis-)similarity matrices for providing the basis for developing disparity measures and essentially improves the accuracy of the subject assignment of cited references (on the basis of the 2nd reference generation) for the measurement of variety but can, on the other hand, also be used to assign the documents themselves to a number of particular subjects. The remaining documents proved to be truly interdisciplinary and require further (qualitative) investigation of knowledge integration. This could be done on the basis of cited literature standing for knowledge that has become integrated into the research in question in conjunction with text analysis using the title, abstracts and keywords or, whenever available, the full texts of the documents. The results of this study form the groundwork for important future tasks: The main future objective is, of course, the creation of measures of variety and disparity with full implementation of the methodology described in the present study. This will be achieved in a separate paper on the different implementations of similarity for disparity and variety measures (Thijs et al., 2021). A secondary task is the individual subject assignment of all papers indexed in the WoS to disciplines according to the ECOOM classification scheme with the possibility of supplementary attribution of IDR labels to documents for policy-relevant applications.

Acknowledgement

The research underlying this study is done within the framework of the project "Interdisciplinarity & Impact" (2019-2023) funded by the Flemish Government. We would like to thank Wouter Jeuris and Lin Zhang for their inspiring discussions.

References

Huutoniemi, K., Klein, J. T., Bruun, H., et al. (2010). Analyzing interdisciplinarity: Typology and indicators. *Research Policy*, 39(1), 79-88.

COSEPUP (2004). *Facilitating interdisciplinary research. National Academies*. "Committee on Facilitating Interdisciplinary Research, Committee on Science, Engineering, and Public Policy", Washington: National Academy Press, p. 2.

Glänzel, W., Schubert, A., & Czerwon, H. J. (1999). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. *Scientometrics*, 44(3), 427–439.

Glänzel, W., & Schubert, A. (2003), A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56(3), 357-367.

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2009). Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, 78(1), 165–188.

Glänzel, W., Thijs, B. & Chi, P.S. (2016). The challenges to expand bibliometric studies from periodical literature to monographic literature with a new data source: The Book Citation Index. *Scientometrics*, 109(3), 2165–2179.

Huang, Y., Glänzel, W., Thijs, B., & Zhang, L. (2021a), *A framework for measuring the knowledge diffusion impact of interdisciplinary research.* Proceedings of the 18th Conference of the International Conference on Scientometrics and Informetrics, Leuven (virtual meeting), 12‑15 July 2021, in this volume.

Huang, Y., Glänzel, W., Thijs, B., Porter, A.L., & Zhang, L. (2021b), *The comparison of various similarity measurement approaches on interdisciplinary indicators*. FEB Research Report MSI_2102, Report No. MSI_2102.

Milojevič, S. (2020), Practical method to reclassify Web of Science articles into unique subject categories and broad disciplines. *Quantitative Science Studies*, 1(1), 183–206.

Porter, A. L., Roessner, D. J., & Heberger, A. E. (2008). How interdisciplinary is a given body of research? *Research Evaluation*, 17(4), 273-282.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(2), 263-287.

Rafols, I. (2014). *Knowledge integration and diffusion: Measures and mapping of diversity and coherence*. In: Ding Y., Rousseau R., Wolfram D. (eds) Measuring scholarly impact (pp. 169-190): Springer, Cham.

Stirling, A. (1994). Diversity and ignorance in electricity supply investment: Addressing the solution rather than the problem. *Energy Policy*, 22(3), 195-216.

Thijs, B., Huang, Y. & Glänzel, W. (2021), *Comparing different implementations of similarity for disparity and variety measures in studies on interdisciplinarity*. Proceedings of the 18th Conference of the International Conference on Scientometrics and Informetrics, Leuven (virtual meeting), 12–15 July 2021, in this volume.

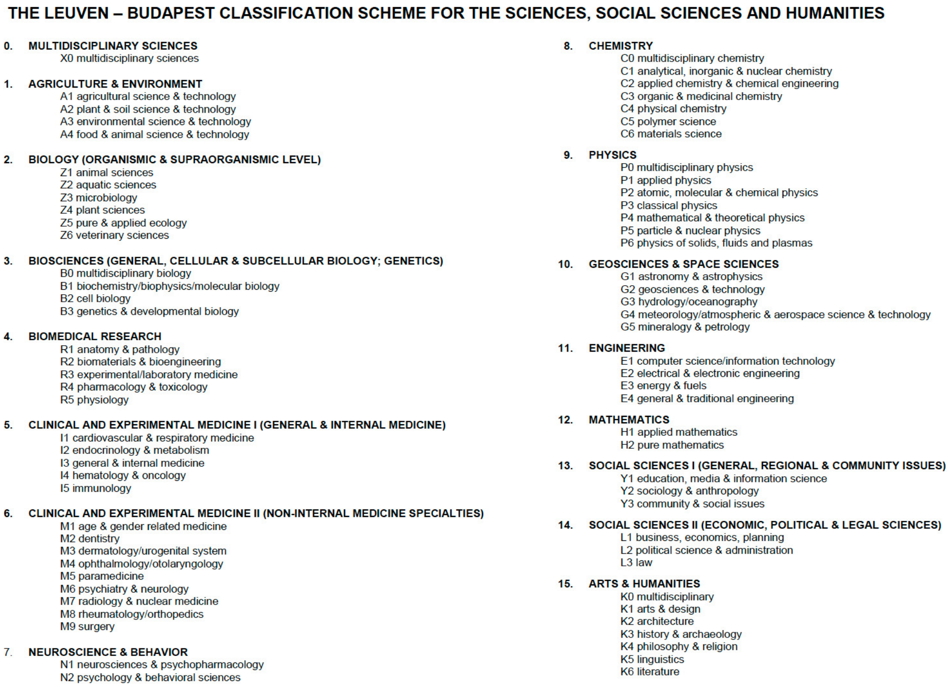
Zhang, L., Janssens, F., Liang, L.M., & Glänzel, W. (2010). Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research. *Scientometrics*, 82(3), 687–706.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: taking similarity between subject fields into account. *Journal of the association for information science and technology*, 67(5), 1257-1265.

Zhang, L., Sun, B., Chinchilla-Rodríguez, Z., Chen, L., & Huang, Y. (2018). Interdisciplinarity and collaboration: on the relationship between disciplinary diversity in departmental affiliations and reference lists. *Scientometrics*, 117(1), 271-291.

Appendix

The revised Leuven-Budapest classification scheme according to Glänzel et al. (2016)



THE LEUVEN – BUDAPEST CLASSIFICATION SCHEME FOR THE SCIENCES, SOCIAL SCIENCES AND HUMANITIES

0. MULTIDISCIPLINARY SCIENCES
X0 multidisciplinary sciences

1. AGRICULTURE & ENVIRONMENT
A1 agricultural science & technology
A2 plant & soil science & technology
A3 environmental science & technology
A4 food & animal science & technology

2. BIOLOGY (ORGANISMIC & SUPRAORGANISMIC LEVEL)
Z1 animal sciences
Z2 aquatic sciences
Z3 microbiology
Z4 plant sciences
Z5 pure & applied ecology
Z6 veterinary sciences

3. BIOSCIENCES (GENERAL, CELLULAR & SUBCELLULAR BIOLOGY; GENETICS)
B0 multidisciplinary biology
B1 biochemistry/biophysics/molecular biology
B2 cell biology
B3 genetics & developmental biology

4. BIOMEDICAL RESEARCH
R1 anatomy & pathology
R2 biomaterials & bioengineering
R3 experimental/laboratory medicine
R4 pharmacology & toxicology
R5 physiology

5. CLINICAL AND EXPERIMENTAL MEDICINE I (GENERAL & INTERNAL MEDICINE)
I1 cardiovascular & respiratory medicine
I2 endocrinology & metabolism
I3 general & internal medicine
I4 hematology & oncology
I5 immunology

6. CLINICAL AND EXPERIMENTAL MEDICINE II (NON-INTERNAL MEDICINE SPECIALTIES)
M1 age & gender related medicine
M2 dentistry
M3 dermatology/urogenital system
M4 ophthalmology/otolaryngology
M5 paramedicine
M6 psychiatry & neurology
M7 radiology & nuclear medicine
M8 rheumatology/orthopedics
M9 surgery

7. NEUROSCIENCE & BEHAVIOR
N1 neurosciences & psychopharmacology
N2 psychology & behavioral sciences

8. CHEMISTRY
C0 multidisciplinary chemistry
C1 analytical, inorganic & nuclear chemistry
C2 applied chemistry & chemical engineering
C3 organic & medicinal chemistry
C4 physical chemistry
C5 polymer science
C6 materials science

9. PHYSICS
P0 multidisciplinary physics
P1 applied physics
P2 atomic, molecular & chemical physics
P3 classical physics
P4 mathematical & theoretical physics
P5 particle & nuclear physics
P6 physics of solids, fluids and plasmas

10. GEOSCIENCES & SPACE SCIENCES
G1 astronomy & astrophysics
G2 geosciences & technology
G3 hydrology/oceanography
G4 meteorology/atmospheric & aerospace science & technology
G5 mineralogy & petrology

11. ENGINEERING
E1 computer science/information technology
E2 electrical & electronic engineering
E3 energy & fuels
E4 general & traditional engineering

12. MATHEMATICS
H1 applied mathematics
H2 pure mathematics

13. SOCIAL SCIENCES I (GENERAL, REGIONAL & COMMUNITY ISSUES)
Y1 education, media & information science
Y2 sociology & anthropology
Y3 community & social issues

14. SOCIAL SCIENCES II (ECONOMIC, POLITICAL & LEGAL SCIENCES)
L1 business, economics, planning
L2 political science & administration
L3 law

15. ARTS & HUMANITIES
K0 multidisciplinary
K1 arts & design
K2 architecture
K3 history & archaeology
K4 philosophy & religion
K5 linguistics
K6 literature

Measuring the excellence contribution at the journal level: An alternative to Garfield's impact factor

Juan Gorraiz [1], Ursula Ulrych[1], Wolfgang Glänzel[2], Wenceslao Arroyo-Machado [3] and Daniel Torres-Salinas [3]

[1] *juan.gorraiz, ursula.ulrych@univie.ac.at*
University of Vienna, Vienna University Library, Dept, of Bibliometrics and Publication Strategies, Boltzmanngasse 5, A-1090 Vienna (Austria)

[2] *wolfgang.glanzel@kuleuven.be*
ECOOM and Dept. MSI, KU Leuven, Naamsestraat 61, Leuven, 3000, Belgium

[3] *torressalinas@go.ugr.es*
Departamento de Información y Comunicación, Universidad de Granada, Granada, Spain

Abstract

The aim of this study is to analyze to which extent the JIF reflects the amount of excellent publications contained in a journal in the corresponding subject category. Therefore, we are introducing two percentile-based indicators in order to measure the excellence contribution at journal level. Calculations of these indicators have been carried out for five different JCR subject categories to investigate the correlation with Garfield's Journal Impact Factor. Differences in the ranking according to all three indicators especially in Quartile 1 of each category are shown and discussed. We have also studied the effect of multidisciplinary journals to the excellence contribution at category level and observed considerable differences between the five categories. In the hard sciences, their omission would lead to neglect a large part of excellent publications. Furthermore, our results hint to the fact that the introduced excellence indicators are very robust considering the types of documents considered for their calculation.

This pilot study shows that the introduction of journal excellence indicators will provide a complete and more accurate picture of the citation impact of a journal than the JIF, because they are informing directly about the total and normalized excellence contribution of each journal to the corresponding subject category.

Introduction

Since its introduction by Garfield in the 1960s (first mention in 1963, Garfield and Sher,1963; Garfield 1972, 1976), the Journal Impact Factor (JIF) is still one of the most common bibliometric indicators when it comes to measuring journal impact (Archambault & Larivière, 2009). Its popularity is unbroken, and not only because its introduction meant a revolution for the scientific community (Lariviere & Sugimoto, 2019). The simple fact that, despite the development of a multitude of new indicators, none of the alternatives has prevailed testifies to the high acceptance of the IF when it is used reasonably (Garfield, 2005; Gorraiz et al., 2020). The past has clearly shown that the JIF is not an all-in-one solution for various issues, which has led to controversial discussions and justified criticism (Todorov & Glänzel, 1988; Moed & van Leeuwen, 1995; Glänzel & Moed, 2002; Moed, 2002; Alberts, 2013; Gorraiz et al., 2012a). In response, several manifestos and statements were published especially due to the increasingly frequent misuse in research-assessment practices (San Francisco Declaration on Research Assessment, ASCB, 2012).

The first edition of the Journal Citation Reports (JCR) – including for the first time the Journal Impact Factor - was launched in 1975 and was based on the fundamental understanding that citations can be used as valuable criterion for the assessment of scientific journals (Garfield, 1976). The more frequently a journal is cited, the higher the recognition of its importance and prestige as information channel in the respective research field.

Researchers started to use the JIF in order to identify adequate publication venues and to optimize their publication strategies. As of its introduction editors and publishers rely on the

JIF in order to estimate the reputation, prestige and market value of their journal portfolio. Furthermore, the JIF opened up a new support tool for librarians to back up decisions about subscriptions, to guarantee the presence of indispensable journals in their collections and to optimize their acquisition strategy. Finally, policy makers have thus gained a quantitative indicator for evaluation purposes, which additionally drove the expansion from its use for scientific information to application in evaluative contexts (cf. Glänzel, 2006).

The JIF has been further developed and improved over the years (extension of the citation window to five years, consideration of the journal self-citations, etc.) and nowadays a number of alternative journal citation-indicators are available such as the h-index for journals (Braun et al., 2006), eigenfactor metrics (Bergstrom et al., 2008; West et al., 2010), SJR (González-Pereira et al., 2010; Guerrero-Bote & Moya- Anegón, 2012), the SNIP indicator (Moed, 2010 a&b) or the CiteScore (cf. van Noorden, 2016). Nevertheless, the new edition of the JCR is eagerly expected each year, which shows the continuing importance of this analytical tool for the scholarly community and for research assessment.

Research assessment exercises are often performed for recent time periods. In these cases, impact analyses relying on citations are not very useful, because in many disciplines the citation window is practically too short for retrieving significant citation numbers.

Although it is not the appropriate indicator to measure the impact of a publication (Waltman & Traag, 2017), the JIF does provide a quick information on the impact and prestige of the journals in which the researcher, group or institution has been able to publish. Being published in journals with high JIF is much more difficult (higher rejection quotes), and successful publication in these journals needs recognition. JIF also helps to identify the top journals in each field according to their impact or prestige. This is why the Journal Impact Factor plays such a key role.

The competition to be included in the Web of science Core Collection and to be indexed as a Q1 journal or to publish in one continues unabated (Osterloh, & Frey, 2014) and is inextricably linked to the question of how the citation impact and prestige of a journal is measured.

However, since the introduction of the IF, many analytical tools have been developed and are available, enabling a very quick and automatic calculation of the percentiles of the most cited publications for each publication year and each subject category (Lozano et al., 2012)..

Nowadays the normalized citation counts like Category Normalized Citation Impact CNCI and the number and percentage of Top 10% and Top 1% most cited publications are essential indicators in citation analyses (Adams et al., 2007; Gorraiz et al., 2012b; Gorraiz & Gumpenberger, 2015).). Top 10% is usually considered as a measure of "excellence".

Therefore, it can be quite interesting to use these normalized indicators as an alternative to the JIF. Does the JIF reflect the amount of excellent publications contained in a journal or in a subject category? Are there other approaches to paint a more precise picture of journal excellence? This is the subject of our study.

Research questions

In order to achieve our objectives to measure the excellence contribution at the journal level, we will answer the following research questions.

1. Can the Journal Impact Factor (JIF) designed to provide a robust and size-independent journal performance measure be supplemented by an indicator of excellence based on the high-end of a journal's publications? Could a proper percentile-based approach result in an improved assessment of the citation impact of a journal?
2. How does the Impact Factor correlate with the proposed percentile-based indicators?
3. Which high-impact journals from the last ten years do not appear in Quartile 1 (Q1/WoS)? What could be the reasons for this?
4. How do multidisciplinary journals affect the indicators?

5. How sensitive are these indicators to the choice of document types, particularly of the so-called ‘citable items’ (i.e., research articles and reviews) instead of all documents types?

Methodology

All documents assigned to the WoS Subject Categories “Virology“ (VIR), “Physics, Condensed Matter” (PHCM), “Economics” (ECO), “Information & Library Science” (ILS) and “History” (HIS) published of the years between 2009 and 2018 were selected and subsequently analyzed in InCites at the journal level.

In this study, we are considering only journals with an Impact Factor, and we are performing the analyses for two different groups: 1) only journals assigned to each WoS subject category according to Journal Citation Reports (“JCR Cat.”), and 2) including all multidisciplinary journals that, according to InCites, have likewise contributed to this category (“JCR Cat. + Multidisciplinary”).

For each journal, we list:

- Number of publications published in this journal in JCR Cat.: p(J)
- Number of excellent publications published in this journal in JCR Cat.: x(J)

For each category we list:

- Total number of publications in JCR Cat: p(T)
- Total number of excellent publications in JCR Cat.: x(T)

In this study the term “excellent publications” or “excellence” is used as synonym for publications belonging to the Top 10% most cited documents in the same JCR Category, publication year and document type.

Beside the Journal Impact Factor retrieved from the JCR Edition 2020, we have calculated the following indicators for each journal:

1. Journal Percentage of Excellent Publications (JPEP) = (x(J)/p(J)) = Number of excellent publications published in this journal in the PY=2009-2018 in this WoS Category / Total number of publications published in this journal in the PY=2009-2018 in this WoS Category
2. Journal Contribution to the Excellence of the Category (JCEC) = (x(J)/x(T)) = Number of excellent publications published in this journal in the PY=2009-2018 in this WoS Category / Total number of excellent publications published in the PY=2009-2018 in this WoS Category.

Both indicators are size dependent: The first one (JPEP) can reach very high values for journals with just few publications in the category, and the second one (JCEC) benefits journals with a large number of publications. Therefore, we have also calculated two further indicators:

3. Journal Brute Excellence (JBE) = JPEP * JCEC = $x^2(J)/(p(J)*x(T))$.
4. Journal Normalized Excellence (JNE) = (x(J)/x(T))/(p(J)/p(T)) = Journal Contribution to the Excellence (JCEC) / Journal Contribution to the Category

The first one reflects the total brute excellence force or brute contribution of the journal to the category. The second one provides the normalized excellence contribution of the journal to the category. Both together provide a more complete picture of the journal excellence.

We are using the JNE especially for the analysis limited to the journals assigned to the JCR Category under study (“JCR Cat.”), because the number of publications of these journals is significant, resulting in relevant JNE values. Note that JNE is inspired by the “Attractivity Index” by Schubert and Braun (1996), which is, in turn, defined based on the model of the Activity Index introduced into Scientometrics by Frame (1977). Both indicators have been used since the late 1980s to reflect a country’s, region’s or other unit’s relative contribution to research productivity and citation impact in given subject fields (cf. Schubert et al., 1989). JNE here expresses a journal’s contribution to the excellence in a given subject. As such JNE, analogously to the above-mentioned indicators by the Hungarian research group, is a balance measure with neutral value 1, i.e. a journal contributes relatively more (less) to the subject's excellence according as JNE > (<) 1. It is not contributing at all, if JNE = 0. The only conceptual deviation of JNE from activity/attractivity is that the balance in not considered across subjects but across units (i.e., journals). A consequence of the "balance" property of this concept is that not all journals can contribute relatively more (less) than expected – some journals assigned to the subject category reflect relatively more excellence than the subject standards, others contribute to subject excellence to a lesser extent.

When analyzing the effect of the multidisciplinary journals, we use the JBE. Multidisciplinary journals contributing rather few publications to the category yield high JNE values, but according to the JBE no significant contributions are achieved.

Pearson Correlations were then performed for the JIF, JPEP, JCEC, JBE and JNE for each of the five categories considering only journals assigned to the subject category ("JCR Cat.") and including also multidisciplinary journals ("JCR Cat. + Multidisciplinary").

In addition, the Gini coefficient has been calculated for all five categories.

Furthermore, we have compared the Q1 journals assigned to each category according JCR (2020) with the Top Journals according to the two new indicators JNE ("JCR Cat."), and JBE ("JCR Cat. + Multidisciplinary").

In order to address research question #4, we have analyzed and discussed the contribution of other journals not directly assigned to the corresponding category, like e.g. the multidisciplinary journals, to the excellence of the category. For this purpose, we have introduced two more indicators:

5. Category Percentage of Multidisciplinarity (CPM) = Number of publications added by multidisciplinary journals not directly assigned to this category according to the JCR (e.g. Nature, Science, PLOS, etc.) / Total number of publications in the category.
6. Category Excellence Degree Multidisciplinarity (CEDM) = Number of excellent publications added by journals not directly assigned to this category according to the JCR (e.g. Nature, Science, PLOS, etc.) / Total number of excellent publications in the category.

Last but not least, we have also performed our analysis not only for the document types articles and reviews, but also for all document types in order to address research question 5.

Results

General Overview

Table 1 gives an overview of the number of journals, publications and excellent publications for each category considered in this study.

Table 1. Overview of the five categories (PY= 2009-2018)

		JCR Category			JCR Cat. + Multidisciplinary			Multidisciplinarity	
Categories	**Document Types**	**Nr Journals**	**Nr Pubs**	**Nr Excellent Pubs**	**Nr Journals**	**Nr Pubs**	**Nr Excellent Pubs**	**Percentage CPM**	**Degree CEDM**
Economy	All types	351	242.035	29987	433	246764	30481	1.92%	1.62%
	Art. / Rev.	343	181852	26332	424	187937	26694	3.24%	1.36%
History	All types	99	82575	8983	126	88258	9902	6.44%	9.28%
	Art. / Rev.	99	25443	7371	119	27992	8157	9.11%	9.64%
Information & Library S	All types	79	92657	7077	144	95329	7213	2.80%	1.89%
	Art. / Rev.	73	33745	5850	131	37828	5950	10.79%	1.68%
Physics High Condensed M	All types	63	285812	29762	100	289678	29934	1.33%	0.58%
	Art. / Rev.	62	278819	29145	98	283149	29307	1.53%	0.55%
Virology	All types	35	83006	7269	145	89138	8248	6.88%	11.87%
	Art. / Rev.	34	65277	6258	140	71280	7174	8.42%	12.77%

Furthermore, it provides information about the differences between document types – all document types (All types) versus Article and Reviews (Art. /Rev.) –, the "Category Percentage of Multidisciplinarity" (CPM) and the "Category Excellence Degree Multidisciplinarity (CEDM)" (see section Methodology). The results show that articles and reviews are mostly responsible for the number of excellent publications in all categories. This is even true for the three categories related to the Social Sciences where big differences between the total number of all document types compared to articles and reviews can be observed.

The lowest percentage of articles and reviews within the excellent publications is observed for "Information & Library Science" (ILS) and "History" (HI) with 82%, followed by "Virology" and "Economics" with around 88% and the highest in "Physics, Condensed Matter" (PCM) with almost 98%. In this study, we are focusing on the document types Articles (Art,) and Reviews (Rev.). In Section 5, the effect of the document types will be further analyzed and discussed.

Table 1 also shows that the category percentage and degree of multidisciplinarity are different according to the subject categories. For the category "Virology" (VIR) it is even more significant when considering the contribution to excellence (CEDM). More than 12% of the excellent publications are published in multidisciplinary journals in the category of "Virology" and around 10% in the category "History". In "Physics, Condensed Matter" (PCM) the effect of the multidisciplinary journals is almost inexistent, and in Economics (ECO) very low. In "Information & Library Science" (ILS) the effect is much higher in the total number of publications (CPM) than in the number of excellent publications (CEDM) as well as for articles and reviews in comparison to all document types.

Showcase Results for the category "Information & Library Science" (ILS)

Table 2 provides an example of the results obtained for the category "Information & Library Science" (ILS) and includes all the indicators mentioned in the methodology.

Table 2. Excerpt of the data retrieved for the category ILS, only Q1 journals according to the JIF 2019 (Article & Reviews, PY=2009-2018)

						Final Indicators	
	Journal Impact Factor	Nr Pubs	Nr Excellent Pubs	% Journal Contribution JPEP	% Category Contribution JCEC	Brute Excellence JBE	Normalized Excellence JNE
INT. JOURNAL OF INFORMATION MANAGEMENT	8,21	796	357	44.85%	6.10%	2.737	2.851
MIS QUARTERLY	5,37	511	333	65.17%	5.69%	3.709	4.143
JOURNAL OF COMPUTER-MEDIATED COMMUNICATION	5,366	338	154	45.56%	2.63%	1.199	2.897
JOURNAL OF STRATEGIC INFORMATION SYSTEMS	5,231	192	79	41.15%	1.35%	0.556	2.616
INFORMATION & MANAGEMENT	5,155	640	294	45.94%	5.03%	2.309	2.921
GOVERNMENT INFORMATION QUARTERLY	5,098	598	231	38.63%	3.95%	1.525	2.456
INFORMATION PROCESSING & MANAGEMENT	4,787	685	151	22.04%	2.58%	0.569	1.401
JOURNAL OF KNOWLEDGE MANAGEMENT	4,745	663	264	39.82%	4.51%	1.797	2.532
JOURNAL OF INFORMETRICS	4,611	734	204	27.79%	3.49%	0.969	1.767
INFORMATION SYSTEMS JOURNAL	4,188	247	83	33.60%	1.42%	0.477	2.136
TELEMATICS AND INFORMATICS	4,139	702	195	27.78%	3.33%	0.926	1.766
JOURNAL O AMERICAN MEDICAL INFORMATICS ASSOCIATION	4,112	1713	525	30.65%	8.97%	2.750	1.948
MIS QUARTERLY EXECUTIVE	4,088	163	32	19.63%	0.55%	0.107	1.248
INT. J. COMPUTER-SUPPORTED COLLABORATIVE LEARNING	4,028	194	55	28.35%	0.94%	0.267	1.802
JOURNAL OF MANAGEMENT INFORMATION SYSTEMS	3,949	406	157	38.67%	2.68%	1.038	2.458
INT. JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE	3,733	1079	222	20.57%	3.79%	0.781	1.308
JOURNAL OF INFORMATION TECHNOLOGY	3,625	217	56	25.81%	0.96%	0.247	1.641
INFORMATION SYSTEMS RESEARCH	3,585	497	204	41.05%	3.49%	1.431	2.610
INFORMATION AND ORGANIZATION	3,3	132	29	21.97%	0.50%	0.109	1.397
JOURNAL OF THE ASSOCIATION FOR INFORMATION SYSTEMS	2,957	317	89	28.08%	1.52%	0.427	1.785
SCIENTOMETRICS	2,867	2921	495	16.95%	8.46%	1.434	1.077

It shows the Q1 journals according to the Journal Impact Factor[1].

Figure 1 shows the correlation between the IF and the two new indicators for all journals of the JCR Category "Information & Library Science" (ILS). The correlation is rather moderate (JBE; r = 0.763, see Table 3), most notably for the normalized JNE (*r* = 0.906, see Table 3), but some of the journals change their position, if a normalized and size-independent indicator (JNE) is used (e.g., JASIST and Scientometrics).

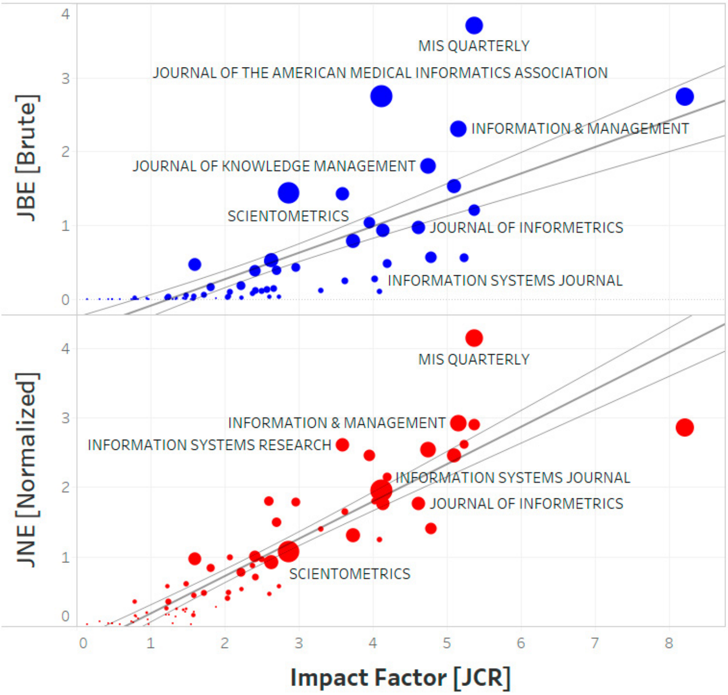


Figure 1. Correlations of the Impact factor with the JBE and JNE in the category Information Library & Science

Table 3 shows the Pearson correlations between all five indicators: JPEP, JCEC, Journal Impact Factor (JIF), JBE and JNE for the category "Information & Library Science" (ILS) for a) only articles and reviews (lower left triangle) b) for all the document types (upper right triangle).

Table 3. Pearson correlations between all measures and indicators for all JCR journals of the category ILS (lower left triangle: Articles & Reviews; upper right triangle: all document types; PY=2009-2018)

A+R vs. all	Pubs	Exc. Pubs	JIF	JPEP	JCEC	JBE	JNE
Pubs	-----	0.077	-0.136	-0.107	0.077	-0.107	-0.029
Exc. Pubs	0.775	-----	0.683	0.696	1.000	0.696	0.890
JIF	0.253	0.700	-----	0.897	0.683	0.897	0.755
JPEP	0.222	0.716	0.906	-----	0.696	1.000	0.830
JCEC	0.775	1.000	0.700	0.716	-----	0.696	0.890
JBE	0.462	0.890	0.763	0.836	0.890	-----	0.830
JNE	0.222	0.716	0.906	1.000	0.716	0.836	-----

[1] The complete dataset will be uploaded in Zenodo.

We will discuss the effects of the differentiation between all document types and articles & reviews later in section 5.

Table 4. Ranking changes for journals of the category ILS according to the JIF in comparison with JBE and/or JNE (green/red= increasing/decreasing positions) (Article & Reviews, PY=2009-2018)

	Ranks			Differences Between rankings	
	JIF	Brute Excellence JBE	Normalized Excellence JBE	JIF and JBE	JIF and JNE
JOURNAL OF HEALTH COMMUNICATION	41	**19**	26	+22	+15
PORTAL-LIBRARIES AND THE ACADEMY	65	**45**	44	+20	+21
JOURNAL OF ACADEMIC LIBRARIANSHIP	55	**38**	45	+17	+10
INFORMATION TECHNOLOGY & MANAGEMENT	56	**40**	36	+16	+20
SCIENTOMETRICS	21	**7**	23	+14	-2
ELECTRONIC LIBRARY	64	**51**	58	+13	+6
ONLINE INFORMATION REVIEW	39	**26**	32	+13	+7
LIBRARY & INFORMATION SCIENCE RESEARCH	47	**35**	35	+12	+12
LIBRARY HI TECH	57	**46**	47	+11	+10
EUROPEAN JOURNAL OF INFORMATION SYSTEMS	27	**17**	13	+10	+14
INFORMATION SYSTEMS RESEARCH	18	**8**	6	+10	+12
JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION	12	**2**	11	+10	+1
ETHICS AND INFORMATION TECHNOLOGY	35	**33**	25	+2	+10
JOURNAL OF INFORMETRICS	9	**11**	15	-2	-6
PROFESIONAL DE LA INFORMACION	44	**49**	57	-5	-13
KNOWLEDGE ORGANIZATION	59	**68**	70	-9	-11
LIBRARY COLLECTIONS ACQUISITIONS & TECHNICAL SERVICES	52	**62**	60	-10	-8
ASLIB JOURNAL OF INFORMATION MANAGEMENT	34	**44**	38	-10	-4
INFORMATION AND ORGANIZATION	19	**30**	20	-11	-1
JOURNAL OF STRATEGIC INFORMATION SYSTEMS	4	**15**	5	-11	-1
REVISTA ESPANOLA DE DOCUMENTACION CIENTIFICA	53	**66**	68	-13	-15
DATA BASE FOR ADVANCES IN INFORMATION SYSTEMS	42	**55**	51	-13	-9
JOURNAL OF ORGANIZATIONAL AND END USER COMPUTING	38	**52**	46	-14	-8
JOURNAL OF GLOBAL INFORMATION TECHNOLOGY MANAGEMENT	45	**60**	56	-15	-11
LEARNED PUBLISHING	26	**42**	41	-16	-15
INFORMATION TECHNOLOGY FOR DEVELOPMENT	22	**39**	37	-17	-15
MIS QUARTERLY EXECUTIVE	13	**32**	22	-19	-9
MALAYSIAN JOURNAL OF LIBRARY & INFORMATION SCIENCE	46	**71**	71	-25	-25

In Table 4, journals in the category ILS are listed. It shows the changes in ranking position, which is traditionally based on the Journal Impact Factor, when applying the Excellence Indicators JBE and JNE. Values in green indicate a higher ranking position compared to the JIF, values in red indicate a lower position.

Portal: *Libraries and the Academy* and *Journal of Health Communication* are the journals that improve their rank position the most due to the excellence indicators. *Malaysian Journal of Library & Information Science* and *Information Technology for Development* are the ones decreasing the most in the brute and normalized excellence rankings.

Comparisons between the five categories analyzed

Table 5 shows the results of the correlation between the Impact Factor and the two Excellence Indicators for all the categories considered in our study: "Information & Library Science" (ILS), “Economics” (ECO), “History” (HIS), “Physics, Condensed Matter” (PHCM) and “Virology“ (VIR).

Table 5. Correlations between JIF, JBE and JNE for the five subject categories analyzed (Article & Reviews, PY=2009-2018)

ILS	JIF	JBE	JNE
JIF	1.000		
JBE	0.763	1.000	
JNE	0.906	0.836	1.000

ECO	JIF	JBE	JNE
JIF	1.000		
JBE	0.544	1.000	
JNE	0.909	0.620	1.000

HIS	JIF	JBE	JNE
JIF	1.000		
JBE	0.624	1.000	
JNE	0.784	0.836	1.000

PHCM	JIF	JBE	JNE
JIF	1.000		
JBE	0.691	1.000	
JNE	0.941	0.799	1.000

VIR	JIF	JBE	JNE
JIF	1.000		
JBE	0.758	1.000	
JNE	0.956	0.867	1.000

The results show that the correlation between the Journal Impact Factor (JIF) and the JNE is higher than between JIF and the JBE. This is expected because JIF and JNE are both size independent.

The correlation between the JIF and the JNE is very high for the JCR Categories “Virology“ (VIR) and “Physics, Condensed Matter” (PHCM) (around 0.95), good for “Economics” (ECO) and “Information & Library Science” (ILS) (around 0.9) and lower for
“History” (HIS) (0.784).

Effect of the multidisciplinarity

Table 6 illustrates strong differences in the effects of the “multidisciplinary journals” in the five selected categories. Categories related to the life sciences and natural sciences show strong influences of such journals compared with the Social Science that are less affected. Of course, we have to keep in mind that humanities and most fields in the social sciences have a lesser weight in the big multidisciplinary journals. In particular, virology (VIR) is the category with the highest presence in multidisciplinary journals. Five multidisciplinary journals are responsible for the largest brute excellence contribution and can be considered as “Q1 journals” in this category according to this indicator. In Economics, four multidisciplinary journals appear among the journals with high contribution to the brute excellence, but not on the top. Also four journals are listed for “Physics, Condensed Matter” (PHCM) with comparably lower ranking positions.

The only multidisciplinary journal ascending to the first quartile in “Information & Library Science” (ILS) is PLOS ONE. As it is well-known, PLOS ONE has a special section for Research assessment and Bibliometrics. However, according to its size, its excellence contribution is not as high as expected (see also Table 1).

Table 6. Effect of the multidisciplinary journals in the JBE Ranking for the journals of the five subject categories (Article & Reviews, PY=2009-2018)

			Contributions		Brute		
Category	Journal	PUBs	JPEP	JCEC	JBE	Rank JBE	Percentile
VIR	PNAS	650	49.38%	4.47%	2.210	4	8
VIR	SCIENCE	118	98.30%	1.61%	1.590	5	10
VIR	NATURE	100	97%	1.35%	1.312	6	12
VIR	NATURE COMMUNICATIONS	176	58.52%	1.43%	0.840	7	14
VIR	NEW ENGLAND JOURNAL OF MEDICINE	15	73.33%	0.15%	0.112	12	24
PHCM	SCIENCE	30	93.33%	0.09%	0.089	11	15
PHCM	NATURE COMMUNICATIONS	178	34.27%	0.20%	0.071	13	18
PHCM	NATURE	26	84.61%	0.07%	0.064	14	20
PHCM	PNAS	84	30.92%	0.08%	0.027	16	23
ILS	PLOS ONE	199	39.19%	1.31%	0.514	17	20
HIS	JOURNAL OF ECONOMIC HISTORY	323	74.30%	2.94%	2.186	2	2
ECO	PNAS	248	56.04%	0.52%	0.292	24	7
ECO	SCIENCE	62	85.48%	0.19%	0.170	37	10
ECO	PLOS ONE	726	11.98%	0.32%	0.039	88	24
ECO	NATURE	10	100%	0.03%	0.037	92	25

Effect of the document types

Finally, we analyzed the effect of considering all types of documents instead of only articles and reviews. As it is common knowledge that there is an asymmetry in the calculation of the Journal Impact Factor. In the numerator, the citations to all types of documents are summed up, while in the denominator only research articles and reviews are considered[2]. In Section 1 we have already analyzed the document types in each category and their contribution to the Excellence (see Table 1). The results corroborate that in the subject categories related to the social sciences (ILS and HIS), other document types than articles and reviews might play a significant role accounting for around 18% of the category excellence.

Furthermore, the two new excellence indicators have been also calculated for all document types and for articles and reviews only (see Table 3). The results underline the role of research articles and reviews in scientific journals. Any reasonable correlation of the number of documents with excellence measure is absent, even slightly negative. Thus it is plausible that the observed Pearson correlation between JIF and JNE is distinctly higher for articles and reviews than for all document types (0.906 versus 0.755), while it is just the opposite for the brute excellence contribution (JBE), where the total number of publication in the category plays a role (0.763 versus 0.897).

Figure 2 shows the correlation of the Journal Impact Factor, and the two excellence indicators (JBE and JNE) for the Q1 journals of the category ILS when considering only articles and reviews (column 2 and 3) and all document types (column 4 and 5), respectively. The results show that, even if the actual indicator values are changing, the distribution of the JBE or JNE as such is not much affected by the considering all document types instead of only 'citable items'. This hints to the fact that our excellence indicators are quite robust or less sensitive to the types of documents considered. In particular, the correlations are very strong, e.g. 0.986 for the Journal Normalized Excellence (JNE), and 0.99 for the Journal Brute

[2] Originally, Garfield used the document types, articles and reviews, also called "citable items" in the JCR Edition. Nowadays, all the proceedings papers published in journals are also considered articles in the Core Collection with the effect of double assignment.

Excellence (JBE), and they corroborate the robustness of both indicators concerning the document types used in their calculation. One possible reason is that the normalizations performed for defining excellent publications are also done by document type (= Top 10% most cited publications of the same document type and publication year in the same category year).

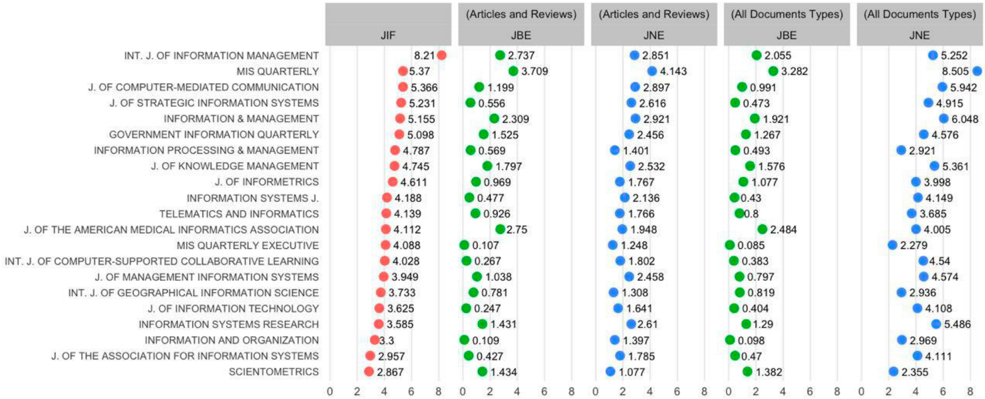


Figure 2. Distribution of JIF, JBE and JNE for all Q1 journals of the category ILS for only Articles and Reviews (2nd and 3rd columns 2 and 3) and all document types (columns 4 and 5).

Conclusions

Due to the precariousness and long half-lives of the citations, the identification of the top journals in each discipline is one of the most requested and used tools in academic evaluation exercises focusing on the assessment of the research performance of the most recent years.

The Journal Impact Factor has established itself as one of the most consolidated instruments for assessing the impact and prestige of the journals where the scientists, research groups, organizations and countries have published in.

To provide a broader view of each journal's contribution to the excellence in each category or field, we have introduced two new indicators, which ideally complement the JIF. The first one, the Journal Normalized Excellence (JNE) measures the normalized excellence contribution of a journal to its subject category. A journal contributes relatively more (less) to the subject's excellence according as JNE > (<) 1.

On the other side, it is also interesting to know the total contribution of a journal to the category excellence, independently of its size. The Journal Brute Excellence (JBE) reflects the total brute excellence force or brute contribution of the journal to the category. It also plays a crucial role in order to estimate the effect of multidisciplinary journals in the category and especially their contribution to the excellence in the category.

This study also reveals that the effect of the multidisciplinary journals is very different according to the category and it is generally stronger in the hard sciences.

In this pilot study, which was restricted to five subject categories, our excellence indicators have shown a robustness concerning the consideration of all types of documents instead of only articles and reviews. Therefore, they provide an amelioration of the inherent asymmetry reflected in the definition and calculation of Garfield's Impact Factor.

Another advantage of our excellence indicators relies on the practical aspect for the measurement of the visibility of publications. When using the Impact Factor for this purpose, there is always a controversial decision: what JCR Edition should be used? There are three possibilities: a) using JIF values of the last JCR-edition for all publications independently of their publication year; b) Using the JCR-edition corresponding to the publication year of each publication; and c) using the mean value of the last *x* years according to the time period under study. Anyone of them is completely satisfactory (Glänzel et al., 2016). Our excellence indicators circumvent this problem because they are based on accumulated measures including the last ten complete publication years and are not restricted to two years or a selected JCR edition.

One of the possible applications of our study is to prevent the use of JCR Categories for the delineation of scientific areas, as has been done in many previous bibliometric studies. Our study warns of serious consequences of this approach, as contributions from multidisciplinary journals are not considered in some categories. For example, reducing the study to only journals of the category in Virology or PHM would mean missing a large part of the scientific breakthroughs and excellent publications, which are regularly published in multidisciplinary journals. In economics, however, this contribution is not notable, and in LIS or HIS, only sporadic contributions can be observed.

Although it is well known that journal impact measures do not work well in the Arts and Humanities and can lead to false interpretations (Repiso et al., 2019), we have also considered the category "History" in an exploratory way. However, in these disciplines it will be crucial to determine which types of publications contribute most to excellence, and this will be part of our future studies.

Future analyses could also be extended to include the effect of interdisciplinarity. Unfortunately, InCites does not offer the possibility to measure this effect, because the subject classification is made on journal level, except for the multidisciplinary journals (on publication level). The recent introduction of the publication based "Citation Topics" may be an improvement in InCites. This topic will also be part of our future analyses.

References

Adams, J., Gurney, K. A., & Marshall, S. (2007). Profiling citation impact: A new methodology. *Scientometrics, 72*, 325–344.

Alberts, B. (2013). Impact factor distortions. Science, 340, 787–787. doi: 10.1126/science.1240319

Archambault, É., & Larivière, V. (2009). History of the journal impact factor: contingencies and consequences. Scientometrics, 79(3), 639-653.

ASCB. (2012). San Francisco Declaration on Research Assessment. Retrieved from: http://www.ascb.org/dora/

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The eigenfactor™ metrics. Journal of Neuroscience, 28(45), 11433-11434.

Braun, T., Glänzel, W. & Schubert, A. (2006). A Hirsch-type index for journals. Scientometrics, 69(1), 169-173.

Frame, J.D., (1977), Mainstream research in Latin America and the Caribbean, *Interciencia*, 2(3), 143-148.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science, 178(4060), 471-479.

Garfield, E. (1976). Preface. In Garfield, E. (Ed.) Journal Citation Reports ® A Bibliometric Analysis of References Processed for the 1974 Science Citation Index ®. Science Citation Index, Volume 9, 1975 Annual.

Garfield, E. (2005). The agony and the ecstasy—the history and meaning of the journal impact factor. *J. Biol. Chem*, *295*, 1-22.

Garfield, E., & Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. American Documentation, 14(3), 195-201.

Glänzel, W., Chi, P.S., Gumpenberger, C., & Gorraiz, J. (2016). Information sources - information targets: evaluative aspects of the scientists' publication strategies. 21st International Conference on Science and Technology Indicators - STI 2016. Book of Proceedings. Woolley, Richard (Ed.). Spain: Editorial Universitat Politecnica de Valencia.

Glänzel, W. (2006). *The 'perspective shift' in bibliometrics and its consequences* (Keynote presentation). First International Conference on Multidisciplinary Information Sciences &

Technologies. PowerPoint presentation available at: http://www.slideshare.net/inscit2006/the-perspective-shift-in-bibliometrics-and-its-consequences.

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. Scientometrics, 53(2), 171– 193.

Gorraiz, J., & Gumpenberger, C. (2015). A flexible bibliometric approach for the assessment of professorial appointments. Scientometrics, 105(3), 1699-1719.

Gorraiz, J., Gumpenberger, C., Schlögl, C., & Wieland, M. (2012a). On the temporal stability of Garfield's Impact Factor and its suitability to identify hot papers. In *Proceedings of STI 2012 Montreal. 17th international conference on science and technology indicators*, Vol 1, pp. 319–332.

Gorraiz, J., Reimann, R., & Gumpenberger, C. (2012b). Key factors and considerations in the assessment of international collaboration: A case study for Austria and six countries. *Scientometrics, 91*(2), 417–433.

Gorraiz, J., Wieland, M., Ulrych, U., & Gumpenberger, C. (2020). De Profundis: A Decade of Bibliometric Services Under Scrutiny. In *Evaluative Informetrics: The Art of Metrics-Based Research Assessment* (pp. 233-260). Springer, Cham.

González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals scientific prestige: The SJR indicator. Journal of Informetrics, 4(3), 379–391. http://dx.doi.org/10.1016/j.joi.2010.03.002

Guerrero-Bote, V.P., & Moya- Anegón, F. (2012). A further step forward in measuring journals' scientific prestige: The SJR2 indicator. Journal of Informetrics, 6, 674-688.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for research metrics. Nature, 520(7548), 429-431.

Lariviere, V., & Sugimoto, C. R. (2019). The journal impact factor: A brief history, critique, and discussion of adverse effects. In *Springer handbook of science and technology indicators* (pp. 3-24). Springer, Cham.

Lozano, G. A., Larivière, V., & Gingras, Y. (2012). The weakening relationship between the impact factor and papers' citations in the digital age. Journal of the Association for Information Science and Technology, 63(11), 2140-2145.

Moed, H.F. (2002). The impact-factors debate: the ISI's uses and limits. Nature, 415, 731-732.

Moed, H. F., & van Leeuwen, T. N. (1995). Impact factors can mislead. Nature, 381(6579), 186.

Moed, H.F. (2010a). Measuring contextual citation impact of scientific journals. Journal of Informetrics, 4, 265-277.

Moed, H. F. (2010b). The source normalized impact per paper is a valid and sophisticated indicator of journal citation impact. *Journal of the American Society for Information Science and Technology, 62*(1), 211–213.

Osterloh, M., & Frey, B. S. (2014). Ranking Games. Evaluation Review, 39(1), 102-129.

Repiso, R., Gumpenberger, C., Wieland, M., & Gorraiz, J. (2019). Impact Measures in the Humanities: A Blessing or A Curse? Book of Abstracts QQML 2019; http://qqml.org/wp-content/uploads/2017/09/Book-of-Abstracts_Final_AfterConf_v1.pdf

Schubert, A., Glänzel, W., & Braun, T. (1989), Scientometric Datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major fields and subfields 1981-1985. *Scientometrics*, 16(1-6), 3-478.

Schubert, A., & Braun, T. (1996), Cross-field normalization of scientometric indicators. *Scientometrics*, 36(3), 311-324.Todorov, R., & Glänzel, W. (1988), Journal citation measures: A concise review. *Journal of Information Science*, 14(1), 47-56.

Van Noorden, R. (2016). Controversial impact factor gets a heavyweight rival. *Nature*. 540(7633), 325–326.

West, J.D., Bergstrom, T.C., & Bergstrom, C.T. (2010). The EigenfactorTM Metrics: a network approach to assessing scholarly journals. College & Research Libraries, 71, 236-244.

Waltman, L., & Traag, V. A. (2017). Use of the journal impact factor for assessing individual articles need not be wrong. arXiv preprint arXiv:1703.02334.

Revisiting the Obsolescence Process of Individual Scientific Publications: Operationalisation and a Preliminary Cross-discipline Exploration

Zhenyu Gou[1], Fan Meng[1], Zaida Chinchilla-Rodríguez[1,2] and Yi Bu[1,3]

[1] *{gouzhenyu, mengfan, buyi}@pku.edu.cn*
Department of Information Management, Peking University, Beijing (China)

[2] *zaida.chinchilla@csic.es*
Consejo Superior de Investigaciones Científicas (CSIC), Instituto de Políticas y Bienes Públicos (IPP), Madrid (Spain)

[3] Center for Complex Networks and Systems Research, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN (U.S.A.)

Abstract

Literature obsolescence has been widely studied in scientometrics. Yet, most existing research has studied a set of publications (e.g., those in a certain discipline or a certain journal; the macro-level analyses) instead of each individual publication (e.g., the micro-level analyses). Among micro-level analyses, Bailón-Moreno et al. (2005) proposed a model by dividing the citation "life cycle" into four periods. Yet, their model remains on a conceptualisation level. To this end, the current paper shows a feasible way of operationalising their model and of implementing a set of cross-discipline publications in the Web of Science database to test its performance. In the operationalisation, we are particularly interested in two stages of literature ageing, namely the sleeping period (SP) and the recognition period (RP). We find that 50% publications in *Arts & Humanities* almost receive no citations in the first five years after their publication; and the distribution of SP and RP varies a lot in different disciplines. Obsolescence differences among publications could shed light on many useful political implications.

Background and Research Objective

Literature obsolescence refers to the phenomenon that the content/value of literatures is increasingly out of date and is utilised less as time goes by (Gupta, 1990). For decades, indicators and mathematical models have been proposed, such as the length of half-life (Tsay, 1998) and the Price index (Egghe, 2010). These pioneer studies have painted a macro-level picture of the publication ageing issue and have shed light on science policy decision-making and library information resource services (Clermont, Krolak & Tunger, 2020; Kinney, 2007; Perrault et al., 1999; Petersen et al., 2014). However, although having shown universal patterns of literature obsolescence, most of the existing studies did not study details of the obsolescence process for *each individual paper*, i.e., a micro-level exploration. We believe that investigating individual paper-level obsolescence helps understand the laws in this citation dynamics more deeply and will guide practices in research evaluation.

As an important contribution to studying individual publications' obsolescence, Bailó n-Moreno et al. (2005) proposed a generalised model for studying individual publications' ageing process (the GMAV model). Their proposed model conceptually partitioned the period after a paper was published into four periods according to its citation rate:

Period X (Period I): the publication has low transition capacity, and it will not be cited;

Period P (Period II): the publication has high transition capacity to influence other literatures;

Period C (Period III): the publication is being cited currently, and it could maintain this period in the future. However, it has a trend of moving to Period N; and

Period N (Period IV): the publication has been cited, but its citation rate has declined currently. It is likely that the publication will transition to Period C, but such a probability will diminish with time.

Nonetheless, Bailón-Moreno and colleagues' work remains at a conceptualisation level without presenting detailed operationalisations, which hinders future bibliometricians from widely adopting their model in practice. To this end, **the research objective of the current paper** is to offer a feasible operationalisation of Bailón-Moreno et al.'s (2005) GMAV model. The operationalisation is based on the annual citation time series of a given publication. We showcase the usage of our operationalisation by particularly focusing on the lengths of Period I (sleeping period, abbreviated as SP) and Periods II+III+IV (recognition period, abbreviated as RP). In the empirical study, we adopt all publications in 1985 in the Web of Science database, which covers papers in all disciplines, and implement a cross-discipline analysis. Our dataset covers at least a 30-year-long time window after they were published to guarantee that the citation window is of sufficient length empirically.

Yet, to quantify the length of Periods I-IV, we also need to empirically define Period V (the period after Period IV), in which publications do not receive citations (no longer being "adopted"). This is necessary when people calculate the length of RP (or Periods II+III+IV). In this paper, our operationalisation also highlights this new contribution.

The rest of this paper is organised as follows. We first present previous related papers on literature obsolescence. We then detail how we operationalise the model. Next, we introduce our dataset and show the empirical results based on all publications and those in different disciplines. We also showcase a mini case study to illustrate more details in our operationalisation. Finally, we discuss potential implications, both methodologically and politically, and suggest future work.

Related Work

Literature obsolescence has been widely studied in scientometrics and bibliometrics since the 1940s (Glänzel & Schoepflin, 1995; Sengupta, 1992). Extant related work can be summarised as two different levels (**macro** and **micro** levels), which differ from each other in terms of their research object – the former focuses on *a set of* publications (e.g., those in a certain discipline) while the latter focuses on individual publications.

On the macro level, half-life is an important indicator for measuring obsolescence. Yet, there are two ways of calculating half-life, namely diachronic and synchronic strategies. Generally, both strategies are equally feasible to measure literature obsolescence, but different results would be obtained with the two strategies when considering other factors such as the growth of literatures (Egghe, Rao, & Rousseau, 1995). The diachronic strategy first selects a set of publications and collect their citations' data in the following several years (after their publication); as for the synchronic strategy, we select a particular year and examines all past publications cited by the publications in this particular year. For example, the half-life proposed by Bernal (1958) is diachronic. Later, another concept of half-life based on the synchronic method was proposed by Burton and Kebler (1960), and this is defined as the publication interval of the 50% recently published literature cited by a certain current publication (of a journal or a discipline), which is also called median citation age. Although different, both strategies have been widely used in the calculation of the degree of literature obsolescence. For example, the Journal Citation Report (JCR) adopts both concepts and offers two indicators for journals, namely cited and citing half-life: the former equals the interval of the 1/2 new citing papers of a journal while the latter is the interval of the 1/2 new references of this journal. While the value of the half-life indicator may vary in different years and disciplines, the average value among years is generally used to describe the degree of obsolescence. In addition to the half-life indicator, people also adopt the Price Index to characterise obsolescence, calculated by dividing the number of references published in the previous five years by the total number of references published in the current year. Egghe (1997) proposed the N-year Price Index to supplement the original Price Index by extending five-year-long time-window of this indicator to any integer, N. He pointed out that there exists a certain relationship between Price Index and half-life indicator. Additionally, on a more theoretical yet crucial level, some research has also been devoted to exploring the law of obsolescence and to constructing some literature obsolescence models (e.g., Avramescu, 2007; Burton & Kebler, 1960).

Compared with macro-level studies, the number of micro-level approaches is much smaller, but it offers more details on the process of publication ageing. For example, Wei and Qian (2005) pointed out that every measurement of obsolescence was an estimation in statistics, and the half-life indicator or Price Index of literature could be regarded as point estimation while micro-level approaches could be seen as interval estimation. In other words, micro-perspective research could calculate the probability that the true value falls into this interval, which outperforms macro-level studies. Zhang and Glänzel (2017) found that it might be extremely different in the distribution of Price Index of individual papers that were published in two journals with a similar Journal Price Index. For example, supposing there are two journals, one is composed of 50% small Price Index papers and 50% large Price Index papers, and the other is completely composed of papers with medium Price Index papers. This would result in a similar macroscopic Price Index (journal's Price Index), yet the microscopic Price Index (paper's Price Index) would be quite different, as aforementioned. As a pioneer, Price (1963, 1976) first linked micro-level studies with the half-life indicator. He extended the concept of half-life, and proposed that the half-life value of an individual publication equals the interval of publication of its new 50% citing papers. Price (1963) indicated that the half-life of a paper was about 1.5 years, which meant that half literatures which cite this literature were published within 1.5 years after the publication of this literature.

Model Operationalisation

In his model, Bailón-Moreno (2005) proposed three transitions based upon the aforementioned four periods: the transitions from period *P* to *C*, from *C* to *N* and from *N* to *C*. These transitions represent an increase or decrease of the literatures' citations in different ageing stages. While Bailón-Moreno's model remains at the conceptualisation level, we operationalise the model in a doable way (and adopt a numbered naming strategy for simplicity):

Period X (Period I): the publication's citation count conforms to a zero-growth model, and the number of citations is equal to or is approximately equal to zero;

Period P (Period II): the publication's citation conforms to an exponential model, and its citation count has an accelerating growth;

Period C (Period III): the publication's citation conforms to a linear model, and its citation count grows at a smooth speed;

Period N (Period IV): the publication's citation conforms to the deceleration stage of a logistic model, and its citation count decreases over the years.

This operationalisation indicates a temporal-based ageing process of an individual paper by examining the temporal change of its annual number of citations. Generally, one paper will have a *Period I* after its publication, then its citations may grow quickly in *Period II*. Later, the growth rate decreases and its citations maintain a slow growth, i.e., *Period III*. It then comes to *Period IV* and decays. Finally, the publication will not be cited and adopted anymore (*Period V*). Note that Period V was not mentioned in Bailón-Moreno et al.'s model. Yet, to operationalise Periods I to IV (especially their lengths), we have to define this period when there are no additional citations received. It is worth noting that some publications will not experience all these periods, e.g., "uncited papers" only have *Period I*. The process is illustrated in Figure 1.

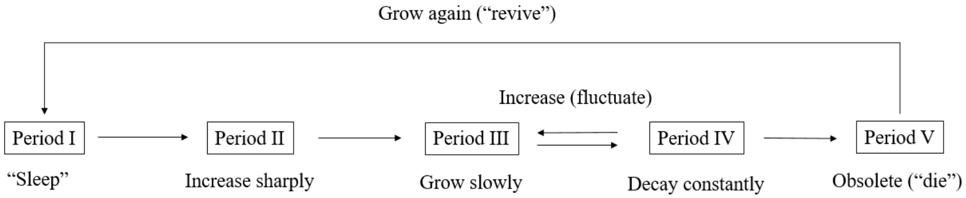


Figure 1. Annotation of the publication obsolescence process.

We then further stipulate the rules of operationalisation as per Raan (2004) by examining the annual number of citations of a publication in each year after it was published:

a) If the number of citations of a specific paper in the first year after its publication is fewer than 2, its state in the first year is *I*; otherwise, the state in the first year is *II*.

b) Given a particular year, if the state of a specific paper in the previous year is *I*, and the number of citations in this year is fewer than 2, the state in this year is *I*; if the number of citations is equal to or greater than 2, the state in this year is *II*.

c) Given a particular year, if the state of a specific paper in the previous year is *II*, and the number of citations in this year is 20% greater than that of previous year, the state is *II* in this year; otherwise, the state is *III*.

d) Given a particular year, if the state of a specific paper in the previous year is *III*, and the number of citations in this year is equal to zero or is 10% smaller than that of previous year, the state in this year is *IV*; otherwise, the state is *III*.

e) Given a particular year, if the state of a specific paper in the previous year is *IV*, and the number of citations in this year is 10% greater than that of previous year, the state in this year is *III*; if the number is not 10% higher than that of previous year and it is fewer than 2, the state in this year is *V*. If either of the two conditions is satisfied, the state in this year is still *IV*.

We adopt iterative check to reduce the volatility as: if the state of the publication in the previous year is the same as that in the following year, the state in the current year would be adjusted to the same period. With this definition, the transitions between states in our new model, shown in Figure 1, and their indications are:

I to II: the literature's citing papers are not subject to their publishing delay (or other unknown factors) and are published successfully, so the literature starts to receive citations.

II to III: the growth of citations has gradually slowed down. For example, prior to this transition, the literature's citations grow rapidly as it is spreading to the author's peers; now, nearly all colleagues are aware of it, so it keeps citation counts at a high level but the growth slows down.

III to IV: the number of citations starts to decrease; knowledge in this literature seems out of date gradually.

IV to III: the transition from Periods IV to III represents the fluctuation of citation counts. For example, the number of citations was decreasing over time, but it increases later as, for instance, some new studies start to discover and cite the literature more frequently.

IV to V: the literature is "dead" and it will not be cited to a large extent.

Experiment

Data

The dataset used in the current study comes from Indiana University Network Science Institute (IUNI) in-house version Web of Science (WoS). It contains the bibliographic data of high-quality journals, conference articles and monographs. Although the WoS is not exempt from limitations on its coverage (Mongeon & Paul-Haus, 2016; Moya et al., 2007), there are at least two reasons why we choose this dataset for our search. On the one hand, the WoS covers a variety of disciplines and a large amount of data at the international level; on the other hand, the journals it covers is selected by experts. We select all WoS papers published in 1985, and set their citation window as 30 years[1].

Table 1. Number of publications in each discipline.

Discipline	# pubs (%)
Arts & Humanities	111,338 (12.46%)
Clinical, Pre-Clinical & Health	253,723 (28.39%)
Engineering & Technology	121,392 (13.58%)
Life Sciences	216,489 (24.23%)
Physical Sciences	179,562 (20.09%)
Social Sciences	100,314 (11.23%)

Specifically, our dataset includes 893,647 distinct publications; these publications have been cited 12,804,905 times. These publications are labelled as *one or more* fields from 250+ subject categories from the classification system of Clarivate[2]. All of these 250+ subject categories are categorised into six disciplines, namely *Arts & Humanities*, *Clinical, Pre-clinical & Health*, *Engineering & Technology*, *Life Sciences, Physical Sciences*, and *Social Sciences*. The numbers of publications under these six disciplines are shown in Table 1. Figure 2 shows the citation distribution of the publications in our dataset.

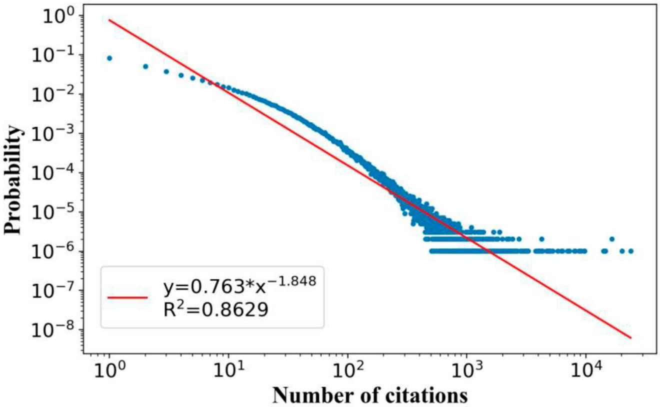


Figure 2. Density distribution of publications' number of citations

Results

For the five-period framework as aforementioned, *Period I* represents the state when a paper has no or a few citations (the annual citation counts should be fewer than 2), and the following three periods (a.k.a., *Periods II+III+IV*) indicate that the paper starts to receive citations, regardless of its trends – increasing or decreasing. To this end, we particularly focus

[1] Some papers were published in January and some papers were published at the end of 1985. Thus, we select the data from 1985 to 2015 (including 2015) to ensure the citation window is not shorter than 30 years, so some papers' citation window will be close to 31 years. But, compared to the citation window, the nuance is so insignificant that we can ignore it.

[2] http://help.prod-incites.com/inCites2Live/indicatorsGroup/aboutHandbook/appendix/mappingTable.html.

on the below two partitions, namely the lengths of *Period I* (sleeping period, annotated as SP) and of *Periods II+III+IV* (recognition period, annotated as RP). Figures 3a and 3b present the distribution of length of SP and RP, respectively.[3]

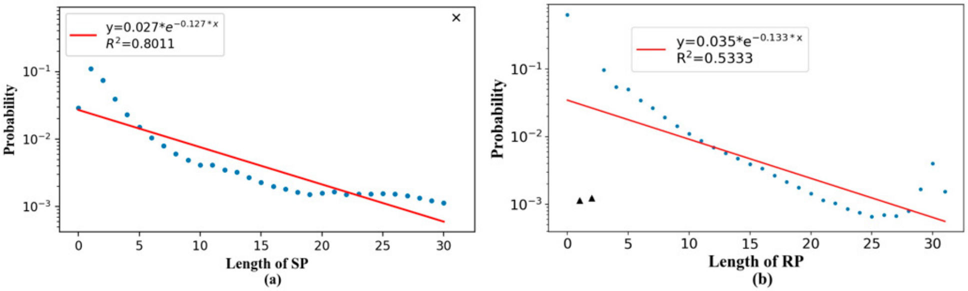


Figure 3. Distribution of the lengths of SP (a) and RP (b).

The black cross ("x") in Figure 3(a) (top-right of the subfigure) represents papers which are always in *Period I* without entering the following period(s). Researchers call them "uncited papers" and they will not be included in the following analyses[4]. The round dots ("·") represent the "cited papers" as they would experience other periods besides *Period I*. By fitting the distribution of SP length for the round dots, we find that the distribution of SP conforms to negative exponential distribution. It suggests that, for the "cited papers", many of them have a short SP, and with prolongation of SP, the number of papers decreases exponentially.

Similarly, we use several functions to fit the distribution of RP and obtain the fitting effect, as Figure 3(b) shows. The distribution of RP also conforms to negative exponential distribution and it shows that most literatures have a short RP. It is worth noting that the fitting effect (R^2) in Figure 3(b) is worse than that in Figure 3(a). By analysing Figure 3(b), we find that the difference mainly comes from the special dots (two black triangles, left-bottom in Figure 3(b)). It can be considered that these triangles represent a special kind of literature called "early rise, rapid decline" (Aversa, 1985) which receives many citations rapidly after its publication and declines quickly after reaching its peak. There are only a limited number of "early rise, rapid decline" papers, and they tend to have a short RP, so the R^2 in Figure 3b is not as good as that in Figure 3(a).

Regarding discipline-wise results, the basic descriptive statistics are shown in Tables 2 and 3. For example, the mean of SP for *Clinical, Pre-Clinical & Health* is 3.98 and that of the RP is 6.96, which illustrates that, on average, the papers belonging to this discipline receive virtually no citations in the first four years after their publication and will tend to be cited in the following seven years (years 5 to 11). After that, the popularity of these papers drops down dramatically and will tend to receive very few citations.

[3] The year of publication is stipulated as 0 instead of 1, and this is also applied for all the following analyses.

[4] The triangle represents the papers that are always in *Period I*; it should be the long-tailed distribution. But we set it as 31 in the X-axis and it is beyond the confines of our citation window (30 years), so our following analyses will not contain these papers (this triangle).

Table 2. SP length of publications in different disciplines.

ID	*Discipline*	*Mean*	*Q1*	*Q2*	*Q3*	*# pubs.*
1	Arts & Humanities	9.92	2	5	19	3952
2	Clinical, Pre-Clinical & Health	3.98	1	2	4	98679
3	Engineering & Technology	5.5	1	3	6	33351
4	Life Sciences	3.62	1	2	4	109039
5	Physical Sciences	3.87	1	2	4	91980
6	Social Sciences	5.98	2	3	7	21310

Note: Publications that only have *Period I* are not included; the same below.

Table 3. RP length of publications in different disciplines.

ID	*Discipline*	*Mean*	*Q1*	*Q2*	*Q3*	*# pubs.*
1	Arts & Humanities	4.66	3	3	5	3952
2	Clinical, Pre-Clinical & Health	6.96	3	5	8	98679
3	Engineering & Technology	6.04	3	4	7	33351
4	Life Sciences	7.36	4	5	9	109039
5	Physical Sciences	6.59	3	5	7	91980
6	Social Sciences	6.79	3	5	8	21310

We can see that some disciplines' SP and RP are similar, such as the *Clinical, Pre-clinical & Health* and *Physical Sciences*, and both their mean values of SP are close to 4. We also observe some differences between the two average values, such as the *Arts & Humanities* (RP mean = 4.66) and *Clinical, Pre-clinical & Health* (RP mean = 6.96). To further explore inter-discipline differences of SP and RP, we employ the Kruskal-Wallis test (KW-test) to test the significance of their difference. KW-test is a nonparametric test for multiple samples. It could be used in the analysis when the distribution of the sample is unknown, so it has a good applicability without limitation of distribution of samples. The distribution of data, SP and RP, does not conform to normal distribution; this is the main reason why we choose the KW-test. SPSS 25.0 is adopted to implement these analyses.

We present the KW-test from two perspectives, namely the whole samples without considering two specific disciplines' statistical differences (Table 4), and for all discipline pairs (Table 5). We can see that there are significant differences in the distributions of SP and RP lengths between disciplines. However, the result in Table 4 is only true for the whole disciplines, and we do not know the significance of difference for two specific disciplines. Therefore, we need to further analyse the test result. The results in Table 5 show that the difference of distribution of SP length is significant and it exists in any two disciplines, which is also the case for the distribution of RP length.

Table 4. KW-test for the whole disciplines. Two stars (**) represent $p < 0.01$.

Test	***Sig.***
KW-test for SP length	.000**
KW-test for RP length	.000**

Table 5. KW-test (*test statistic*) results of SP (bottom-left) and RP (top-right) lengths for discipline pairs.

Discipline	AH	CPH	ET	LS	PS	SS
AH	-	-66341.7**	-42062.5**	-73832.9**	-57364.4**	-52464.3**
CPH	65338.7**	-	24279.2**	-7491.2**	8977.3**	13877.4**
ET	42155.7**	-23183.1**	-	-31770.5**	-15302.0**	-10401.8**
LS	80215.1**	14876.4**	38059.5**	-	16468.5**	21368.6**
PS	74864.1**	9525.4**	32708.4**	-5351.0**	-	4900.1**
SS	26098.9**	-39239.9**	-16056.8**	-54116.3**	-48765.2**	-

Note: (1) Significance: *: p<0.05; **: p<0.01. (2) Abbreviations: Arts & Humanities (AH); Clinical, Pre-Clinical & Health (CPH); Engineering & Technology (ET); Life Sciences (LS); Physical Sciences (PS); and Social Sciences (SS).

Based on these analyses, we find that the mean of SP length of *Arts & Humanities* is the longest and its mean of RP length is the shortest. In contrast, the mean of SP length of *Life Sciences* is the shortest and its mean of RP length is the longest. There exist significant differences in the distribution of SP and RP lengths between any two disciplines, which indicates that the differences of obsolescence processes between science disciplines exist objectively.

A case study

To depict more nuances for the SP and RP of literatures, we select several cases according to the lengths of SP and RP from *information science & library science*. The first case (Paper A) has a short SP and RP. The second case (Paper B) has a short SP but a long RP. The third case (Paper C) has a long SP and RP. As the paper with a long SP but a short RP is so special, it needs to have some characteristics of two kinds of literatures – the literature we called "sleeping beauties" (van Raan, 2004) and the literature we called "early rise, rapid decline" (Aversa, 1985). As this kind of paper is too rare to carry out a case study, the analysis in this subsection does not include such a case. The bibliographic information of the cases is shown in Table 6. Their numbers of citations in each year are illustrated in Figure 4. Again, the citation window is from 1985 to 2015.

Table 6. Bibliographic information of the three cases. All three papers were published in 1985.

	Title	***Author***	***# cits***	***# refs***	***Journal***
A	Critical Thresholds in Co-Citation Graphs	Shaw, W. M.	18	30	*Journal of the American Society for Information Science*
B	Alternative Measures of System Effectiveness - Associations and Implications	Srinivasan, A.	218	29	*MIS Quarterly*
C	Finding Minimal Enclosing Boxes	O'Rourke, J.	41	11	*International Journal of Computer & Information Sciences*

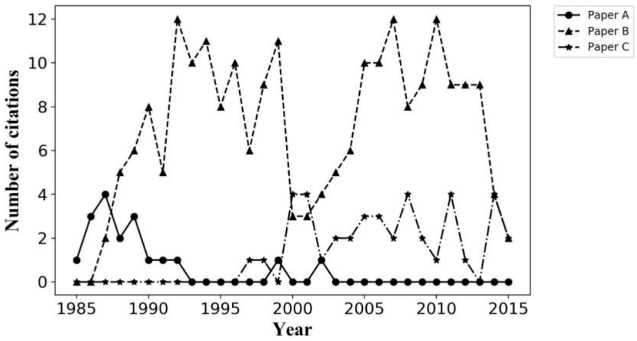


Figure 4. Annual number of citations over the years for the three selected cases.

Paper A (Shaw, 1985) discussed the threshold of co-citation graphs. We can see from Figure 4 that this publication only received one citation exactly in the year when it was published. In the second year after publication (i.e., 1987), it had a short RP and reached its citation peak. After that, its number of citations decreased, and around 1991, it had been obsolescent and had no or few citations in the following two decades. By analysing those citing papers of Paper A, we find that the first citation of Paper A is a self-citation and that its first non-self-citation appeared in 1986.

Paper B (Srinivasan, 1985) studied the effectiveness of Management Information Systems (MIS). Compared with Paper A, Paper B's topic is more concerned with many active scholars – the author researched users' behaviours and system utility and proposed a standard for system evaluation. This research was quite innovative, which was reflected in its early citing papers. Some scholars (Zahedi, 1987) mainly focused on the difference between Paper B and previous empirical studies, while other scholars (Montazemi, 1988) mainly concentrated on the measurement indicators and methods mentioned in Paper B. The great contribution of this publication might be the reason for its long RP. Furthermore, the citing papers of Paper B might improve its popularity further. The mean number of citations of its first 10 citing papers is 845 as there is a "super" citing paper with 7849 citations. Even if that paper is excluded, the mean of citation counts of its first nine papers is 60. These citing papers may improve the citation-based impact of Paper B implicitly.

Paper C studied the minimal volume box for a set of 3-D points (O'Rourke, 1985). In this publication, the author proposed a new algorithm which could find all minimal volume boxes for a set of n-points with the complexity of $O(n^3)$. From Figure 4, we can see that Paper C had a long SP of 15 years. However, the academic contribution of Paper C exists objectively, which is reflected by its long RP. Thus, its long SP might be affected by some external factors, such as the author's career stage and the research topics.

Discussion

This paper operationalises a conceptual model of studying the obsolescence of scientific publications. The operationalisation has three **highlights**. Firstly, we operationalise the model by purely using the annual number of citations of publications, which is feasible for many future bibliometricians. Secondly, the operationalisation contains a new period (Period V) compared with Bailón-Moreno et al.'s (2005) model. The new period captures the "death" state of a publication, indicating that the article will not be cited (to a large extent). The new period helps quantify the length of Period IV in a more accurate way. Finally, our study emphasizes a complex feature of citation dynamics and provides empirical evidences against the use of short-term citation on metrics in research assessment exercises.

We do not quantify the length of Periods I to IV one by one in our empirical study. More particularly, we are interested in two periods in publications' citation life cycle, namely the sleeping period (SP) and recognition period (RP). We find a significant difference regarding the ageing patterns in various domains. For example, we see that 50% of publications belonging to *Clinical, Pre-Clinical & Health* have no or very few citations (annual citation counts equal 0 or 1) during the first two years after their publication (see Table 2). In addition, the case study indicates that there are many complex, external factors (e.g., self-citations, author's career stage, etc.) that may affect SP and RP lengths.

For future bibliometricians, there are two ways of applying our operationalisation. On the one hand, as shown in the current paper, they can purely characterise the lengths of SP (Period I) and RP (Periods II+III+IV) as an easier strategy; this strategy particularly functions when the number of focal publications is great. On the other hand, as citations are at the basis of several quantitative measures increasingly used in evaluation criteria aimed at evaluating career trajectories of scholars (Edwards & Roy, 2017) and research performance of institutions (Bornmann, Haunschild, & Mutzm, 2020), bibliometricians and research evaluators could quantify the lengths of each individual period one by one and present a more nuanced picture of each obsolescence process. That will show more accurate details in research evaluation.

Regardless of which strategy is utilised, our operationalisation offers many interesting and useful implications for science policy makers and funding providers. From tables 2 and 3, we observe quite different patterns for various domains. For example, *Arts & Humanities* has a mean value of 9.92 in terms of SP (~2.5 times that of *Physical Sciences*) but a mean value of 4.66 for RP (the shortest). This indicates to funding providers that they should establish a different rule/criterion to evaluate scholars/works in *Arts & Humanities*. Specifically, these works should be given a longer time to be discovered compared with other domains; more detailedly, the mid-term examination of scholars in this discipline may need to be softened.

Limitations and future work

This pilot study has many limitations. From the perspective of operationalisation, the lengths of Periods II, III, IV, and V are not considered separately in our empirical study. We, therefore, are going to implement more detailed analyses of these periods with more sophisticated strategies, such as advanced time series analysis and parameter settings in operationalisation. Additionally, we observe a special but crucial phenomenon from the citation curves of the publications – the "revival" – and these special publications will experience another *Period II* after *Period V*. This indicates that, although a certain publication has been "dead" (receiving no citations) for quite a long period, sometime it is cited again by a very recent publication. In the future, we are also going to explore this phenomenon by extending the current operationalisation framework.

Acknowledgements

The authors are grateful to all members in the KD lab at Peking University and for the financial support from the National Social Science Foundation (No. 20BTQ054). The suggestions from reviewers and conference organisers are also helpful to polish the current paper. This research was supported – in part – by the Lilly Endowment Inc. through its support for the Indiana University Pervasive Technology Institute, and – in part – by the Indiana METACyt Initiative. The Indiana METACyt Initiative at Indiana University was also supported – in part – by the Lilly Endowment Inc.

References

Aversa, E. S. (1985). Citation patterns of highly cited papers and their relationship to literature aging: a study of the working literature. *Scientometrics, 7*(3-6), 383-389.

Avramescu, A. (2007). Actuality and obsolescence of scientific literature. *Journal of the American Society for Information Science, 30*(5), 296–303.

Bailón-Moreno, R., Jurado-Alameda, E., Ruiz-Baños, R., & Courtial, J. P. (2005). The unified scientometric model. Fractality and transfractality. *Scientometrics, 63*(2), 231-257.

Bernal, J. D. (1958, November). The transmission of scientific information: a user's analysis. In *Proceedings of the International Conference on Scientific Information* (Vol. 1, No. 960, pp. 77-95).

Bornmann, L., Haunschild, R., & Mutz, R. (2020). Should citations be field-normalized in evaluative bibliometrics? An empirical analysis based on propensity score matching. *Journal of Informetrics, 14*(4), 101098.

Burton, R. E., & Kebler, R. W. (1960). The "half-life" of some scientific and technical literatures. *American Documentation, 11*(1), 18–22.

Clermont, M., Krolak, J., & Tunger, D. (2020). Does the citation period have any effect on the informative value of selected citation indicators in research evaluations? *Scientometrics*. https://doi.org/10.1007/s11192-020-03782-1

Price, D. J. D. S. (1963). *Little science, big science*. New York: Columbia University Press.

Price, D. J. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5), 292–306

Edwards, M. A., & Roy, S. (2017). Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental engineering science, 34*(1), 51-61.

Egghe, L., Rao, I., & Rousseau, R. (1995). On the influence of production on utilization functions: obsolescence or increased use?. *Scientometrics, 34*(2), 285-315.

Egghe, L. (1997). Price index and its relation to the mean and median reference age. *Journal of the American Society for Information Science, 48*(6), 564-573.

Egghe, L. (2010). A model showing the increase in time of the average and median reference age and the decrease in time of the Price Index. *Scientometrics, 82*(2), 243-248.

Glänzel, W., & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature. *Journal of information Science, 21*(1), 37-53.

Gupta, U. (1990), Obsolescence of physics literature: Exponential decrease of the density of citations to Physical Review articles with age. *Journal of the American Society for Information Science, 41,* 282-287.

Kinney, A. L. (2007) National scientific facilities and their science impact on nonbiomedical research. *Proceedings of the National Academy of Sciences of the United States of America, 104*(46), 17943–17947

Mongeon, P., and Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics, 106* (1), 213–228.

Montazemi, A. R. (1988). Factors affecting information satisfaction in the context of the small business environment. *MIS Quarterly, 12*(2), 239.

Moya-Anegón, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., González-Molina, A., Muñoz-Fernández, F. J., González-Molina, A., & Herrero-Solana, V. (2007). Coverage analysis of SCOPUS: A journal metric approach. *Scientometrics, 73* (1), 57–58.

O'Rourke, J. (1985). Finding minimal enclosing boxes. *International Journal of Computer & Information Sciences, 14*(3), 183–199.

Perrault, A. H., Madaus, R., Armbrister, A., Dixon, J., & Smith, R. (1999). The effects of high median age on currency of resources in community college library collections. *College & Research Libraries, 60*(4), 316-339.

Petersen, A. M., Fortunato, S., Pan, R. K., Kaski, K., Penner, O., Rungi, A., …, & Pammolli, F. (2014) Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences of the United States of America, 111*(43), 15316–15321.

Sengupta, I. N. (1992). Bibliometrics, informetrics, scientometrics and librametrics: An overview. *Libri, 42*(2), 75.

Shaw, W. M. (1985). Critical thresholds in co-citation graphs. *Journal of the American Society for Information Science, 36*(1), 38–43.

Srinivasan, A. (1985). Alternative measures of system effectiveness: Associations and Implications. *MIS Quarterly, 9*(3), 243.

Tsay, M. Y. (1998). Library journal use and citation half-life in medical science. *Journal of the American Society for Information Science, 49*(14), 1283-1292.

Van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics, 59*(3), 467–472.

Wei, X. X, & Qian, J. (2005). Micro measurement method and interval estimation of literature aging indicator. *Journal of Intelligence, 24*(8), 31-32. (in Chinese).

Zahedi, F. (1987). Reliability of information systems Based on the critical success factors—Formulation. *MIS Quarterly, 11*(2), 187.

Zhang, L., & Glänzel, W. (2017). A citation-based cross-disciplinary study on literature aging: part I—the synchronous approach. *Scientometrics, 111*(3), 1573-1589.

Peer Review versus Bibliometrics. Do the Two Methods Produce Identical Results in Measuring Academic Reputation?

Katerina Guba[1], Mikhail Sokolov[2] and Angelika Tsivinskaya[3]

[1] *kguba@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

[2] *msokolov@eu.spb.ru*
European University at St. Petersburg, Sociology Department, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

[3]*atsivinskaya@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

Abstract

The purpose of this study is to assess the validity of citation metrics based on disciplinary representative surveys. The present project fills the gap by comparing citation rankings for individual scientists with expert judgements collected through a survey of 3,689 Russian sociologists. We used scientometric indicators not limited only to traditional metrics (Hirsch index and Impact Factor), but also special indicators developed by the Russian Index of Science Citation. We conclude that scientometric indicators predict the names of the most influential sociologists with significant accuracy. Citation metrics and survey results converge to a large extent, at least with regard to the goal of identifying the disciplinary elite. We found two types of deviations—false positive and false negative identifications. False positive identifications (high results in scientometrics, but low in the survey) are explained by specialization in writing textbooks and the explicit gaming of metrics, which can, however, be identified.

Introduction

The issue of validity of citation metrics was raised with the launching of the first citation index by Eugene Garfield in the early 1960s. The idea behind the Science Citation Index was to create the infrastructure to enable effective searching of scientific information. However, citation metrics had begun to be considered as a way to measure academic recognition very soon after the launching of the index (Cole & Cole, 1967). The interest in citation rankings was boosted by the expansion of quantitative indicators for research assessment of individual academics, departments, and disciplines. Many countries have launched performance-based research funding systems in which citation metrics are an integral part (Hicks, 2012). Besides the use of scientomentics at the national level, the phenomenon of 'citizen bibliometrics' has come into existence—citation metrics are used by groups, other than professional, including scientists and administrators in everyday evaluation of individuals practitioners (Hammarfelt & Rushforth, 2017). The reliance on quantitative indicators of research productivity was perhaps the most important recent transformation of academic institutions (Auranen & Nieminen, 2010; Espeland, Sauder & Espeland, 2016; Muller, 2018). As a result, the issue of the validity of citation rankings has been discussed widely based on empirical attempts to demonstrate whether citation data are a valid means of measuring scientific performance.

One way to assess the validity of metrics is by studying the degree of their correspondence with expert judgements (So, 1998; Serenko & Dohan, 2011). Researchers conduct academic reputation surveys, in which specialists in a discipline evaluate journals, institutions, individuals, or texts. Disciplinary surveys are considered a reliable way to measure academic quality—recognition or impact—directly, rather than by a proxy through citations. A review of various studies on the validity of citation metrics demonstrates that not much research has been conducted with a focus on individual scholars. While this can be explained by the fact that individual scholars are seldom evaluated in the course of national evaluations (So, 1998), they are nevertheless assessed in the context of faculty hiring, tenure, or promotion. In addition, establishing validity using survey data is less common due to the complexity of organizing a large-scale disciplinary survey. Our research strives to solve the classic problem of assessing the validity of citation metrics by comparing citation rankings for individual scientists and expert judgements collected through a survey of Russian sociologists.

Literature review

The most common method of estimating the validity of citation rankings is to compare expert judgements and citation metrics; high correlation suggests that the validity of bibliometrics has been established (So, 1998). This validation relies on the idea that peer review is the gold standard for scientist evaluation (Wainer & Vieira, 2013; Aksnes & Taxt, 2004). Peer reviewers are considered to be capable of conducting both objective and nuanced evaluation. The advantage of peer-based rankings is their suitability for the development of national and regional rankings for which citation data are not always relevant (Mahmood, 2017) as global citation databases represent mainly English-language journals (Mongeon & Paul-Hus, 2016). During recent decades a number of studies have examined how citation indicators can be correlated with the results of peer reviews. The studies focus on different types of research outputs from entire nations down to individual researchers.

The first way to obtain peer review results is to use evaluations conducted in the course of national evaluation exercises. Governments provide data to researchers to assess the correspondence between peer review scores and citation metrics. Aksnes and Taxt (2004) examined how the outcome of the evaluations conducted by the Research Council of Norway correlated with various citation metrics. In a study of the natural sciences, they found weak positive correlations for all indicators: for the three citation-based metrics the correlation coefficient varied between 0.24 and 0.46 (Aksnes & Taxt, 2004). Wainer and Vieira (2013) conducted research to determine the correlation of various bibliometric measures to peer judgements of scientists working in 55 different scientific fields. They used the evaluation system in Brazil as their source of data. Peers evaluated the scholarship level for individual scholars, and the change in a scientist's level might be considered as the result of the peer evaluation.

The second method of gathering peer judgements is conducting surveys among active researchers in the field for their opinion on the importance of individuals, departments, or journals. More recent studies attempt to establish the validity of citation rankings using journals as an object. Serenko and Dohan (2011) examined whether expert judgements and journal impact measures generate consistent ranking lists in the field of artificial intelligence. Ellis and Durden (1991) studied journals in the field of economics and found that journal reputation is a reflection of the scientific impact measured through citation metrics, although for some journals the correspondence is not large. They found that highly ranked journals tend to have better reputations than their citation impact would predict, and that lower-ranked journals tend to have poorer reputations than their impact would justify (Ellis & Durden, 1991).

So (1998) noted that comparisons between citation metrics and peer evaluation seldom rely on individual scholars as a source of data, compared to journals, research articles, or institutions. We review further the main studies that use individuals as the subject to compare citation metrics and expert opinions. Norris and Oppenheim (2010) conducted a survey of library and information science academics to examine how the results of peer review can be correlated with the h-index. The survey was designed to collect data on the perceived quality of the world's leading academics. The result was a correlation between 0.397 and 0.518 for the h-index, depending on the citation database, which is considered to be medium or large effects (Norris & Oppenheim, 2010). Li et al. (2010) extended the analysis of Norris and Oppenheim (2010) by examining how the peer judgments correlate with citation metrics from different databases, not only from WoS. They confirmed the results of the previous study by discovering strong significant correlations of citation metrics with the expert judgments of 42 LIS experts. However, they were cautious of using citation-based indicators as a cost-effective alternative to the use of experts as the strongest correlation was only 0.552.

Roettger (1978) pioneered an attempt to determine the elite system of a discipline over time. He found agreement on the top ten contributors in political science, but there is less agreement down the ranking as well as diminishing agreement over time. Robey (2012) compared the reputational ranking of political scientists with the ranking by citation and found that the reputational method produces a different list of prominent scholars from the results of citation metrics. He concludes that the reputational method probably includes a number of ways to contribute to the field, while the citation method identifies a more specific type of contribution. So (1998) compared the citation method with peer review judgements in the evaluation of individual scholars in the field of communication by asking respondents to rank the 10 selected scholars in their respected field. He intended to find out how the size of the research specialty and scholar prominence matter for these evaluation methods. So (1998) found results suggesting that the citation method and the peer review provide basically the same outcomes: only a few evaluated scholars are "misplaced" in the peer review method. He also found that correspondence between citation evaluation and peer review is more stable for smaller research specialties which provide more stable results. So (1998) demonstrates the validity of the citation method and concludes that citation data might be used as an alternative and even a substitute for peer review in evaluation of research performance. All these studies point to the necessity of conducting surveys on large samples that guarantee a range of research interests.

To conclude, numerous studies demonstrate mixed results: "whereas several investigations conclude that both methods may be used as substitutes, others report negligible or even negative correlations between the obtained journal scores" (Serenko & Dohan, 2011). Wainer and Vieira (2013) found that the strength of the effect depends on the academic discipline, which prevents generalizing from previous research in peer judgement correlations in one discipline to other research fields. It is also important to conduct studies for each national context.

Methods

The Russian Index of Science Citation (RISC), launched in 2005, is integrated with a full-text platform, the Scientific Electronic Library, which indexes more periodicals than RISC (Moskaleva et al., 2018). RISC provides opportunities for both scientometric research and disciplinary survey. First, in contrast to international citation databases, RISC provides access to citation data based on a wide range of Russian sources, which makes it possible to study non-Western Russian social sciences and humanities. Second, RISC has developed a convenient storage system for scientometric information, which allows extracting citation metrics from profiles for a large sample of scholars. Third, RISC has elaborated a number of additional indicators that are not available in international citation databases. These indicators have been implemented to indirectly indicate the quality of the research and, in some cases, the degree of citation gaming: (1) the share of self-citations, (2) the share of publications and citations from the most selective journals, (3) and (4) the impact factors of the publications in which the papers are published and cited, and (5) the share of citations by co-authors.

The main method for collecting empirical information is the online reputation survey of Russian sociologists for evaluating the degree of convergence of expert judgments and metrics. Forming a sample of scientists is especially problematic for a fragmented discipline that is divided into practically isolated research groups or political camps. The sample should take into account the variety of disciplines and, ideally, cover all scientists. We believe that RISC reduces the problem of selecting scientists to participate in the survey as it is difficult for active scholars to bypass the RISC-indexed journals. Recently, Russian universities were required to register their entire faculty on RISC to provide information for state reports. Many universities created a special position that was responsible for the input of information into the database. Thus, we can conclude that RISC contains a relatively representative dataset of Russian scholars. Therefore, RISC covers a population of scientists close to the general population of the scientific field.

We selected authors (a) who have at least three articles in RISC over the past five years (since we strive to survey active scientists); (b) most of whose articles were published in journals classified by RISC as belonging to sociology, or most of whose citations come from such sources; (c) who are registered in the RISC system and who provided an email address in their profiles. Overall, we ended up with 3,689 profiles to whom the survey was sent by email. Subsequently, we received 818 responses (over 20%) which can be considered a good response rate for online surveys. The respondents accurately represented the general population in terms of the organizations with which the scholars are affiliated.

Results

To measure academic reputation, we developed several questions: (1) who among the Russian sociologists [they] would suggest to include in the national jury of experts which evaluates the work of colleagues; (2) who among the Russian sociologists [they] would nominate for an honorary medal for an important contribution to the development of sociology in Russia; (3) who has published research in recent years that [the respondent] found to be particularly interesting (not necessarily in [the respondent's] area of specialization, and (4) who has published research in recent years that [the respondent] found to be particularly interesting in [the respondent's] area of specialization. The sequence of questions may have influenced the results.

One would expect the answers to provide a different set of sociologists. The medal suggests lifelong achievement and the answer to this question will be a list of recognized scholars, while the recent publication of important articles will yield more names of young scientists. We also expected that the distribution of the authors of the influential works should be the least concentrated, since the respondents themselves work in different fields. The distribution of the twenty most frequently mentioned figures in each of the nominations shows that the lists of names obtained from answers of different questions were largely the same. In total, 10 people were on all four lists, another 10 in three out of four lists, six in two lists, and 12 in only one list, and in the vast majority of cases, these 12 share the last lines with someone. In three of the four lists, the largest number of nominations refers to the same three names. There are only two cases when a top 10 scholar in one list fell below the top 20 in at least one of three other lists. We conclude that academic status in sociology is a generalized category, and reputation is less specific than one might expect. So for further calculations, we add the number of times a sociologist was mentioned in any of the four lists of nominations. It provides us with a variable with a larger number of observations.

The data obtained require the use of specialized regression techniques that take into account the nature of such data (negative binomial regression and zero-inflated models). We used negative binomial regression (detailed models can be provided by request). First, we entered the main indicators separately controlled by the research field (Models 1–6). Model 7 presents the main indicators all together. The next model is the one with all the main indicators supplemented with additional more sophisticated indicators (Model 8). For each model, we consider four metrics: AIC, BIC, and two characteristics of the sorting accuracy—the proportion of predicted values based on the scientometric indicators among the top 20 and top 100. The AIC and BIC metrics can be interpreted as the accuracy with which the model approximates the distribution over all values of the dependent variable, and top points out its ability to predict the position at the top of the rating.

All the models offer a significant improvement over the basic model with only controls (having an AIC and BIC of about 13,000). Moreover, for models with a single main variable (Models 1–6), the best results are obtained with the number of articles in the RISC Core (most selective journals) and with the Hirsch index. Models with all the main indicators (Model 7) and with additional indicators (Model 8) produce even better results. The full model gives 63% predictions for the top 100 (with 10 in the top 20); however, one must bear in mind the danger of overfitting the model. In addition, we present how survey and scientometric rankings are matched with each other (Table 1).

Table 1. Match between reputational and citation results (by scholars).

		Reputational ranking			
		Top-20	Top-50	Top-100	Top-200
Citation ranking	Top-20	11	13	13	13
	Top-50	15	21	31	36
	Top-100	17	31	50	59
	Top-200	19	42	69	89

How can the observed mismatches be explained? We distinguished two types of deviant cases: false positive and false negative. False positive cases arise if we do not find the expected effect, i.e., people with high citation rates who do not have high scores in the survey, while false negative cases, in contrast, are sociologists with a low number of citations and a considerable number of nominations. In general, we found 7 cases of people included in the top 20 by citation metrics who are located below the third hundred by survey results. Among false positive scholars, we find a distinct group of those who demonstrate critical values across the entire set of additional indicators: (1) a high level of self-citation (20–40%) and citations by co-authors (40–70%), (2) a low impact factor of publishers and citing journals (0.25–0.35), (3) a very low number of publications and citations in selective journals. Taken separately, these observations can have a completely different meaning, and, as a consequence, the corresponding variables are not significant.

Conclusion

We used scientometric indicators not limited only to traditional metrics, but also special indicators that were developed by RISC (and have no analogues in other citation databases) and are better able to indicate the quality of scientific work. The results of the study allowed us to conclude that scientometric indicators predict the names of the most influential sociologists with significant accuracy. Citation metrics and survey results converge to a large extent, at least with regard to the goal of identifying the disciplinary elite. Contrary to popular expectations, there is no clear evidence that the presence of support groups results in individuals not favored by most peers scoring high in citation metrics or in a survey. We found two types of deviations—false positive and false negative identifications. False positive identifications (high results in scientometrics, low in the survey) are explained by specialization in writing textbooks and explicit gaming of metrics, which can, however, be identified. False negative identifications (low in scientometrics, high in the survey) are less pronounced and are mainly associated with age or an elite narrow specialization (for example, in methodology).

Acknowledgments

The paper was prepared with the support of Russian Science Foundation No 21-18-00519.

References

Aksnes, D. W., Langfeldt, L. & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, 9 (1).

Auranen, O. & Nieminen, M. (2010). University research funding and publication performance — An international comparison. *Research Policy*, 39 (6), 822–834.

Bornmann, L. & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64 (1), 45–80.

Cole S. & Cole J. (1967). Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review*, 32 (3), 377-390.

Ellis, L. V. & Durden, G. C. (1991). Why economists rank their journals the way they do. *Journal of Economics and Business,* 43 (3), 265–270.

Espeland, W. N. & Sauder, M. (2016). *Engines of anxiety: Academic rankings, reputation, and accountability*. Russell Sage Foundation.

Hammarfelt,B. & Rushforth, A. D. (2017). Indicators as judgment devices. An empirical study of citizen bibliometrics in research evaluation. *Research Evaluation,* 26 (3), 169–180.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41(2), 251-261.

Li, J., Sanderson, M., Willett, P., Norris, M. & Oppenheim, C. (2010). Ranking of library and information science researchers. Comparison of data sources for correlating citation data, and expert judgments. *Journal of Informetrics,* 4 (4), 554–563.

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton University Press.

Mahmood, K. (2017). Correlation between perception-based journal rankings and the Journal Impact Factor (JIF). A systematic review and meta-analysis. *Serials Review,* 43(2), 120–129.

Mongeon, P. & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics,* 106(1), 213–228.

Moskaleva, O., Pislyakov, V., Sterligov, I., Akoev, M. & Shabanova, S. (2018). Russian Index of Science Citation. Overview and Review. *Scientometrics*, 116(1). 449-462.

Norris, M. & Oppenheim C. (2010). Peer review and the h-index: Two studies. *Journal of Informetrics,* 4 (3), 221–232.

Robey, J. (1982). Reputations vs citations. Who are the top scholars in political science? *APSC*, 15, 199–200.

Roettger, W. (1978). Strata and stability. Reputations of American political scientists. *APSC,* 11, 6–12.

So, C. Y. K. (1998). Citation ranking versus expert judgment in evaluating communication scholars. Effects of research specialty size and individual prominence. *Scientometrics,* 41 (3), 325–333.

Serenko, A. & Dohan, M. (2011): Comparing the expert survey and citation impact journal ranking methods. Example from the field of Artificial Intelligence. *Journal of Informetrics*, 5 (4), 629–648.

Tahamtan, I. & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics*, 12 (1), 203–216.

Wainer, J. & Vieira, P. (2013). Correlations between bibliometrics and peer evaluation for all disciplines. The evaluation of Brazilian scientists. *Scientometrics,* 96 (2), 395–410.

Use of Scientometrics in Evaluation of Grant Proposals

Katerina Guba[1], Alexey Zheleznov[2], Elena Chechik[3] and Angelika Tsivinskaya[4]

[1] *kguba@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

[2] *azheleznov@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

[3]*echechik@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

[4]*atsivinskaya@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, Gagarinskaya st. 6/1 A, St. Petersburg (Russia)

Abstract

We study the effect of using quantitative indicators in grant allocation by using the natural experiment with the increasing publication threshold for principal investigators between two waves of grant competitions. The policy change provides us with the opportunity to study whether reliance on bibliometric indicators brings better results in the grant evaluation process. The following questions are explored in greater detail: whether the policy affected the distribution of funds to researchers with a better publication record, the strategies of increasing publications by individual researchers, and the differences, if any, in policy effects between disciplines. We found that the selection among physicists in the first period was already effective as the grant recipients are prolific authors who publish a lot of highly cited papers. Although social scientists demonstrated a relatively weak ability to publish internationally, the increase in scientometric expectations has improved the publication record regarding the quantity and quality of publications.

Introduction

Researchers recognize that grant funding now plays a more significant role in supporting science compared to institutional funding (Auranen & Nieminen, 2010; Grimpe, 2012; Maisano et al., 2020; Wang et al., 2020). Grant funding is expected to trigger competition among scientists, leading to more visible academic results (Grimpe, 2012; Maisano et al., 2020). The prevalence of grant funding raises the issue of its effectiveness. Are grants an effective mechanism for achieving high academic results? Empirical research has mainly answered this question by analyzing the results of supported projects. The analysis shows that grants do not always increase scientific research effectiveness, and if they do, the effect is relatively small. The results require a shift in focus from analyzing project outcomes to analyzing project selection procedures. Can the process for evaluating applications be improved so that grants are awarded to more capable scientists? What role should the scientometric indicators of researchers play in the selection procedure? We will answer these questions by analyzing data on project leaders' scientific performance who received financial support from the Russian Science Foundation.

The role of objective indicators in selecting applications as opposed to expert assessments is widely discussed in the scientific literature. Research demonstrates that the chances of achieving results are lower for those who did not perform well before receiving the grant (Fedderke & Goldschmidt, 2015). If scientometric indicators make it possible to predict the project's subsequent effectiveness, then evaluation procedure may rely on citation metrics.

Additional empirical evidence is required to confirm this thesis, given the policy implications. However, in general, research focuses more on analyzing the results of grants rather than the achievements of grant recipients. The data on the Russian scientists-winners of the Russian Science Foundation competitions will make it possible to contribute to the discussion about the role of bibliometric indicators in the evaluation of grant proposals.

The Russian case is especially indicative given that scientometric methods are actively applied to assess the research performance. To participate in the grant competition, a scientist is required to meet the publication threshold. Moreover, over the past five years, the thresholds for publication of project managers have increased remarkably. If for the 2014 competition, the project leader in natural sciences obliged to have at least three articles in indexed journals published in three years, in 2017, they required five papers in five years. For social and humanitarian scientists, the changes turned out to be more dramatic. Although the number of publications remained the same, the local journals, not indexed in international databases, disappeared from the list of approved sources.

This change in the policy provides us with the opportunity to study whether reliance on bibliometric indicators brings better results in the grant evaluation process. We focus on recipients of research funding from the Russian Science Foundation, which outlined its policy that established a threshold in the number of indexed publications for a principal investigator. The following topics are explored in greater detail: whether the policy affected the distribution of funds to researchers with a good publication record, the strategies of increasing publications by individual researchers, and the differences, if any, in policy effects between disciplines. We propose that social sciences might experience a less positive effect from increasing publication threshold given that quantitative indicators are considered to have less validity in measuring academic quality. Our analysis bases not only on the analysis of the winners' academic performance but also on comparing the publications of the leaders with a sample of highly cited articles in the relevant fields.

Literature review

Evaluation of the effectiveness of grant funding is dominated by studies that analyze the results of supported research projects, as measured by the quality and quantity of publications. Empirical research differs in data collection strategy - researchers either analyze various foundations' cases, tracking publications that resulted in supported projects. In studying individual foundations' performance, a common approach is to compare the performance of scholars who have received grants with a group of similar authors without a grant.

Empirical studies on individual funds show contradictory results. For example, a study of Italian scientists (Maisano et al., 2020) demonstrates that grant recipients did not have greater productivity than those who did not receive the grant. Analysis of publications by grant applicants in New Zealand showed an increase in publications by 3-15% and citations by 5-26%, depending on the type of financial support (Gush et al., 2018). According to the data on the Turkish foundation's effectiveness, grant funding may not have an exact effect on a significant increase in the number of publications and their subsequent citation (Tonta, 2018). The same result was obtained by evaluating young scientists' performance awarded by Russian grants (Saygitov, 2014).

Thus, studies show that the grant funding does not always bring effective results if we expect that supported projects produce high-impact publications (Fedderke & Goldschmidt, 2015). Even if researchers find an effect, it is often relatively small (Morrillo, 2019; Wang & Shapira, 2015). This may be due to the peculiarities of specific funds, which differ in selection procedures and general policies. For example, a comparative study found that Chinese foundations lead to a high number of publications resulting from grants. In contrast, EU grants are less effective, which can be explained by their emphasis on the project's social impact (Wang et al., 2020). Besides, large grants have significant transaction costs and, therefore, not as effective as one might expect (Clark & Liorens, 2012). On the other hand, it is possible that the selection procedures are not meritocratic - the funding does not go to the best scientists. This assumption has been tested in several studies.

Meritocracy in the procedure means that resources go to those who have already shown high achievements (Maisano et al., 2020). Researchers (Fedderke & Goldschmidt, 2015) collected data on the scientific achievements of those who received additional financial resources and those who did not receive them. The comparison showed that funding led to increased performance; however, the increase in publications is very modest depending on scientists' previous achievements. Grimpe (2012) showed that even the most significant competitions fail to attract the best scientists. An analytical report (Technopolis, 2008) presents a large-scale study of principal investigators of large European grants. In particular, grant recipients are compared to academics in similar fields who have not received grants. In general, it is shown that grant recipients are 2-3 times more likely to be authors of the most highly cited articles (10% of managers are the author of at least one article from the pool of the most highly cited publications (1%); 37% have at least one publication from the pool of 10% most cited articles).

In this regard, it is believed that it is better to strive to concentrate resources in the hands of highly productive scientists because it is then that a high return on investments in science can be expected (Maisano et al., 2020). It may be worth formally starting to rely on quantitative performance indicators when deciding which project should receive funding (Fedderke & Goldschmidt, 2015). We propose to focus on the analysis of the previous scientific performance of scientific leaders and to analyze the increasing requirements for scientometric indicators. We include in the analysis the differences by field of science.

Methods

The data collection process began by producing a list of all principal investigators who had received research funding for projects from the Russian Science Foundation in 2014 and 2017. The Russian Science Foundation was selected for its policy to establish a publication threshold for a principal investigator. These publications are prerequisite for the application process. After producing a list of principal investigators, the next step in data collection was gathering data on publications of these investigators in journals listed under Scopus. Publications were matched with the bibliometric characteristics of journals where articles were published. Scopus was selected due to its coverage. It is important to pay attention to the fact that the calculation of research output of Russian institutions significantly depends on the database used. A study of Russian publication demonstrates that the share of papers from Russia indexed in Scopus is higher in Scopus than it is in WoS. For example, Sterligov expressed the view that WoS sells “top quality,” while Scopus sells “top scope,” explaining why the number of journals indexed in Scopus is almost twice as high as it is in the WoS (Sterligiv, 2017). Although the increase in the share of Russian papers during 2012–2016 occurs in both databases, there is a difference in the main factor. For both Scopus and WoS, the coverage of proceedings was expanded, but in the case of Scopus, the increased number of publications also reflects the inclusion of more Russian language journals (Moed, Markusova & Akoev, 2018). The observational window is five years before the year of application.

Results

Our primary empirical strategy is to compare two waives of grant recipients: the first group of scientists received their grant in 2014 before the scientometric threshold was established, while the second group won a competition in 2017 after the new rules were applied. Further, we briefly report our main results.

In the second wave, 92.7% of scientists who received a grant in social sciences had published a piece in the journals indexed by Scopus, while in the first wave, there were only 54.3% of them. For physicists, this percentage is consistently high - 100% and 96.3% in the first and second periods, respectively.

Table 1 presents the number of publications that grantees published in five years before they receive a grant. On average, before receiving a grant in physics, scientists published 42 articles in the first wave and 35 articles in the second wave. At the same time, the median value does not change - 25 papers for both periods. We also do not see a significant change in the average number of publications among social scientists. On average, social scientists publish five articles.

Table 1. Descriptive statistics for number of publications of grantees.

Field	*Time period*	*N grants*	*mean*	*min*	*max*	*var*	*med*
Physics	First wave	115	42	2	466	4158	25
	Second wave	82	35	2	304	1551	25
Social science	First wave	94	5	1	29	25	3
	Second wave	41	5	1	27	21	4

Table 2 shows that the new research leaders had published papers in more selective set of journals (according to CiteScore metrics). For example, in the first wave, the social sciences investigators had articles published mainly in Q4 (59%). In the second period, the share of Q4 decreased to 29.6% - most of the lost Q4 percent was taken by Q1, Q2 and Q3. Physicists also experienced an "overflow" from lower quartiles to higher ones. Results are significant for both research fields (for physics Pearson chi2 – 69.4263, for social science – 42.9478, p< 0.0001). This can be explained both by an increase in the journals' quartiles and by a better selection of grant recipients in favor of those who publish in selective journals.

Table 2. Publications by quartiles.

	Physics		*Social science*	
Journal quartiles	*First wave (%)*	*Second wave (%)*	*First Wave (%)*	*Second wave (%)*
Q1	2653 (53.7)	1437 (52.4)	30 (12.3)	50 (24.3)
Q2	710 (14.9)	452 (16.5)	13 (5.4)	30 (14.6)
Q3	800 (16.7)	600 (21.9)	57 (23.4)	65 (31.5)
Q4	703 (14.7)	253 (9.2)	144 (59)	61 (29.6)
Total	4776 (100)	2742(100)	244 (100)	206 (100)

We also found that CiteScore and Percentile are higher for physicists than for social scientists. However, at the same time for social science, we see more pronounced positive dynamics for Percentile - in the first period Percentile reached on average 33, in the second - 44 (an increase of 33%). Physicists also grew in Percentile from 59 in the first period to 64 in the second (an increase of 10%).

Given the fact that the number of Russian journals in Scopus almost doubled for analyzed periods, we compared not only the total number of publications, but the share of publications in Russian journals. In the second period, the share of international publications grows for both fields. For the principal investigators in physics in the first period, 73% of publications are non-local, and in the second period is 80%. For recipients of social sciences grants, the growth is even higher - from 41 to 59%.

We further compared the grant recipients with the Russian scientists who authored the most cited papers published in two time periods. We stored 5% of Scopus's most cited articles, where authors / co-authors are scientists with Russian affiliation. To collect the sample of highly-cited publications of Russian authors, we queried in Scopus all publications with Russian affiliation separately for physics and social science, and for both periods. Then, we sorted all publications by the number of citations and restricted the sample to 5 percent of most cited publications. Our aim was to find whether scientists among the authors of these highly-cited articles are from the list of grant recipients.

Table 3 shows that the share of investigators who authored highly cited publications in physics is higher than in social science. However, in social science, we see positive dynamics for two periods. Thus in the first period, 10.6% of leaders had published highly cited papers, in the second already 17.1%. At the same time, we see fewer grants were awarded in the second period in both areas. With a decrease in the number of grants, we conclude that the second wave brings an increase in the selected applications' quality in social sciences.

Table 3. Share of grant recipients who author highly cited papers.

Field	*Time period*	*N grant recipients with highly cited publications*	*N highly cited papers authored by grant recipients*	*Share of grant recipients authored highly cited papers*
Physics	First wave	82	504	71.3
	Second wave	56	270	68.3
Social science	First wave	10	13	10.6
	Second wave	7	14	17.1

To conclude, our results demonstrate the relatively positive effect of reliance on bibliometric indicators in the grant evaluation process. We found that the selection among physicists in the first period was already effective as the grant recipients became prolific authors who publish a lot of highly cited papers. In the second wave, the reduction in the number of grants led to a decrease in the names we looked for in highly cited articles, which is why we found fewer articles for physics. In contrast, the social science investigators are less prolific in publishing internationally. Usually they publish fewer papers in the most selective set of journals that can be explained by the local traditions and limited source coverage in international databases. At the same time, social scientists have improved the publication record regarding the quantity and quality of publications. The improvement in performance in the social sciences and humanities can be attributed to at least two reasons. Scientists could correctly perceive the incentives to publish in international journals and start publishing in prestigious journals, which they had not practiced before. The other explanation is that the competition began to be acquired by scientists who had previously published good articles but rarely participated and/or won in competitions. The information about the organization distribution rather speaks in favor of the second explanation. Social scientists affiliated with the Russian Academy of Sciences or a leading university began to win the competition more often - in 2014, there were 65% of them, in 2017, they were already 82%. The rotation of the competition winners took place in favor of scientists from leading organizations with stronger publications, which should rather be assessed as a positive effect.

References

Auranen, O. & Nieminen, M. (2010). University research funding and publication performance—An international comparison. *Research Policy*, 39 (6), 822-834.

Clark, B.Y. & Lliorens, J. J. (2012). Investments in scientific research: Examining the funding threshold effects on scientific collaboration and variation by academic discipline. *Policy Studies Journal*, 44 (4), 698-729.

Fedderke, J. W. & Goldschmidt, M. (2015). Does massive funding support of researchers work?: Evaluating the impact of the South African research chair funding initiative. *Research Policy,* 44 (2), 467-482.

Grimpe, C. (2012). Extramural research grants and scientists' funding strategies: Beggars cannot be choosers? *Research Policy,* 41 (8), 1448-1460.

Gush, J., Jaffe, A., Larsen, V. & Laws, A. (2018). The effect of public funding on research output: the New Zealand Marsden Fund. *New Zealand Economic Papers*, *52*(2), 227-248.

Maisano, D. A., Mastrogiacomo, L. & Franceschini, F. (2020). Short-term effects of non-competitive funding to single academic researchers. *Scientometrics,* 123 (3), 1261-1280.

Moed, H., Markusova, V. & Akoev, M. (2018). Trends in Russian research output indexed in Scopus and Web of Science. *Scientometrics*, 116, 1153-1180.

Saygitov, R.T. (2014) The impact of funding through the RF President's Grants for Young Scientists (the field – Medicine) on research productivity: A quasi-experimental study and a brief systematic review. *PLoS ONE*, 9(1): e86969.

Sterligov, I. (2017). The monster ten you have never heard of: Top Russian scholarly megajournals', *Higher Education in Russia and Beyond*, 1, 11.

Technopolis. (2010). *Bibliometric profiling of Framework Programme participants*. Technopolis Group, Brighton.

Tonta, Y. (2018). Does monetary support increase the number of scientific papers? An interrupted time series analysis. *Journal of Data and Information Science*, 3(1), 19-39.

Wang, J. & Shapira, P. (2015) Is there a relationship between research sponsorship and publication impact? An analysis of funding acknowledgments in nanotechnology papers. *PloS One*, 10, e0117727.

Wang, L., Wang, X., Piro, F. N., & Philipsen, N. J. (2020). The effect of competitive public funding on scientific output: A comparison between China and the EU. *Research Evaluation*, rvaa023, https://doi.org/10.1093/reseval/rvaa023

Early detection of problems with scientific papers using Twitter data: A case study of three retracted COVID-19/SARS-CoV-2 papers

Robin Haunschild [1] and Lutz Bornmann [2]

[1] *R.Haunschild@fkf.mpg.de*
Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart, Germany

[2] *bornmann@gv.mpg.de*
Science Policy and Strategy Department, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany

Abstract

Methodological mistakes, data errors, and scientific misconduct are considered prevalent problems in science that are often difficult to detect. In this study, we explore the potential of Twitter for discovering these problems with publications. We analyzed tweet texts of three retracted publications about COVID-19 (Coronavirus disease 2019)/SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) and their retraction notices. We did not find early warning signs in tweet texts regarding one publication, but we did find tweets that casted doubt on the validity of the two other publications shortly after their publication date. An extension of our current work might lead to an early warning system (based on Twitter data) that makes the scientific community aware of problems with certain publications. Other sources, such as blogs or post-publication peer review sites, could be included in such an early warning system.

Introduction

Certain manuscripts are retracted after publication, although they have been reviewed by colleagues in journal peer review (Sotudeh, Barahmand, Yousefi, & Yaghtin, 2020). There are various reasons for retractions of publications. The main reasons are data errors, methodological mistakes, and, or scientific misconduct. These errors can further be "distinguished between 'reputable' errors (those arising despite the investigators' best efforts to avoid them) and 'disreputable' errors (those arising from disregard of acceptable scientific practice)" (Zuckerman, 2020, p. 954). Fang, Steen, and Casadevall (2012) reported that "2,047 biomedical and life-science research articles indexed by PubMed as retracted on May 3, 2012 revealed that only 21.3% of retractions were attributable to error. In contrast, 67.4% of retractions were attributable to misconduct, including fraud or suspected fraud (43.4%), duplicate publication (14.2%), and plagiarism (9.8%)."

In cases of minor problems with a published paper, a correction is usually published. However, if correcting the problems alters conclusions of the study, retraction can be the only possibility for correcting the scientific record. Retraction decisions are usually not done with levity because retractions may damage the scientific reputation of authors and journals. Retractions may cast a poor light on the research capacity of authors and the functionality of journal peer review. Therefore, one should expect that retractions do not occur very often. However, occurrences of scientific misconduct seem to become known to the scientific public often enough so that the topic of scientific misconduct has become "a thriving area of research" (Zuckerman, 2020, p. 945). The Web of Science (WoS; Birkle, Pendlebury, Schnell, & Adams, 2020) has a document type named "retraction" that has indexed documents with publication years since 1974.

Rumors about data errors, methodological mistakes, and scientific misconduct can spread very quickly on social media such as Twitter (a popular micro-blogging platform, see https://www.twitter.com) (Teixeira da Silva Jaime & Dobránszki, 2019). Sugawara et al. (2017) found that discussions on scientific misconduct regarding a study on stimulus-triggered acquisition of pluripotency (STAP) cells in Japan in 2014 occurred on Twitter before they surfaced in newspapers. Do Twitter users mention problems with papers (early) that are later retracted? Are Twitter users who mention retracted publications informed about the paper status? Do they explicitly mention the issues why the publication has been retracted? These are the questions we try to answer in this case study by analyzing the tweets that mention three different retracted publications about COVID-19 (Coronavirus disease 2019)/SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) and their retraction notices. Since Twitter is a very fast medium, tweets might be used as early indicator for problems with certain publications.

We selected the three publications (Bae et al. (2020a), Wang et al. (2020a), and Mehra, Desai, Ruschitzka, and Patel (2020)) from a post from Retraction Watch about retracted COVID-19/SARS-CoV-2 papers (https://retractionwatch.com/retracted-coronavirus-covid-19-papers/). The three publications have received considerable attention in online media as can be seen from their Altmetric Attention Scores.

Methods

We used the Altmetric.com application programming interface (API) for extracting tweet identifiers for any tweets that mentioned either any of the three retracted publications or their retraction notices on Twitter. The tweet texts and other meta data of the tweets that mentioned any of the publications or retractions on Twitter were downloaded from the Twitter API. We connected to both APIs using R (R Core Team, 2019) with the R packages httr (Wickham, 2017a) and RCurl (Lang & the CRAN team, 2018). The tweets were downloaded between the 6th and the 9th of August 2020 via the Twitter API using R (R Core Team, 2019) and stored in local SQLite database files using the R package RSQLite (Müller, Wickham, James, & Falcon, 2017). Functions from the R package DBI were used for sending database queries (R Special Interest Group on Databases (R-SIG-DB), Wickham, & Müller, 2018).

The R package ggplot2 was used for plotting the time evolution of tweets (Wickham, 2016). The R package tidyverse (Wickham, 2017b) was used for analyses of the Twitter user profiles. The R packages tm (Feinerer, Hornik, & Meyer, 2008), NLP (Hornik, 2014), SnowballC (Bouchet-Valat, 2014), wordcloud (Fellows, 2014), and RColorBrewer (Neuwirth, 2014) were used for producing word clouds. Before searching in tweet texts and displaying them as word clouds, the tweet texts were converted into lower case characters and only alpha-numeric characters were kept. English stop words, punctuation characters, and needless whitespaces were removed.

We downloaded 42,746 tweets for Mehra, Desai, et al. (2020) and 67,038 tweets for its retraction notice, 10,875 tweets for Bae, et al. (2020a) and 3095 tweets for its retraction notice, and 5428 tweets for Wang, et al. (2020a) and 53 tweets for its retraction notice. The retraction notices have received very different numbers of tweets compared to the corresponding papers.

In order to receive information about the people mentioning the three publications or their retraction notices in tweets, we used the classification scheme proposed by Toupin, Millerand, and Larivière (2019) and a modified version of the R code provided by Toupin (2020). The general idea behind most classifications is to capture the interest of a particular account to share scholarly papers using self-descriptions in the profiles (see Haustein, 2019; Toupin & Haustein, 2018; Vainio & Holmberg, 2017).

The results of the classification show that personal accounts are the largest single group for all documents except Wang et al. (2020b). In the case of Wang, et al. (2020b), the user group 'faculty and students' is the largest single group. The user groups 'bots' and 'journals and publishers' are the least frequently occurring groups for all six documents. We expect the most frequent hints to problems with papers especially from the user group 'faculty and students' since this group includes people with the necessary expertise. A critical discussion of papers can be scarcely expected from the user groups 'bots' and 'journals and publishers'.

Results

Twitter activity of Bae, et al. (2020a) and Bae et al. (2020b)

Bae, et al. (2020a) studied the effectiveness of surgical and cotton masks in blocking SARS–CoV-2. The study was published on 6 April 2020 and retracted on 2 June 2020 because they "had not fully recognized the concept of limit of detection (LOD) of the in-house reverse transcriptase polymerase chain reaction used in the study" (Bae, et al., 2020b). In this case, the retraction was made because of a methodological error that was not detected in the peer review process. Figure 1 shows the tweets per day that mentioned either the paper or its retraction. The publication dates of the paper and retraction notice are marked as gray vertical lines. Most tweets mentioning the publication occurred before retraction. However, the second-most frequent tweets per day were recorded on the day of retraction. Therefore, the tweet texts are analyzed using two different time periods: (i) before the publication date of the retraction and (ii) since the publication date of the retraction.

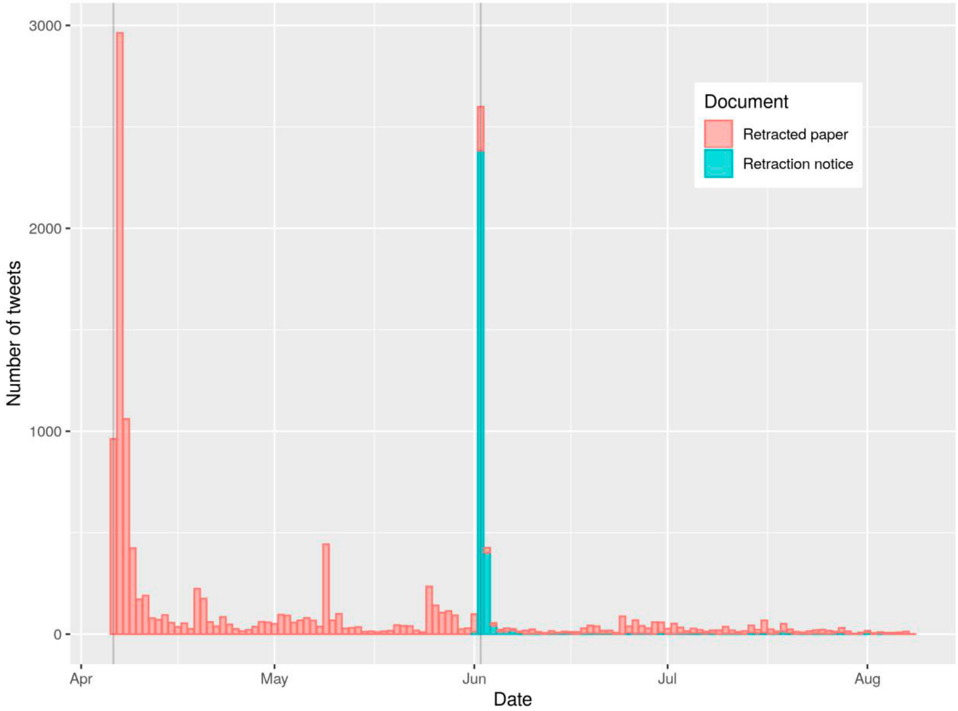


Figure 1: Tweets per day that mentioned either the paper Bae, et al. (2020a) or its retraction Bae, et al. (2020b). The publication dates of the paper and the retraction notice are marked as gray vertical lines.

Figure 2a shows a word cloud from tweet texts based on the tweets mentioning Bae, et al. (2020a) before publication date of the retraction. Figure 2b shows a word cloud from tweet texts based on the tweets mentioning Bae, et al. (2020a) or Bae, et al. (2020b) since publication date of the retraction. In both word clouds, the term 'mask' is the most prominent one. The terms 'surgical' and 'cotton' are the second- and third-most prominent ones in Figure 2a. The two terms are much less prominent in Figure 2b although all three terms ('masks', 'cotton', and 'surgical') appear in the title of both publications.

The publication was retracted because a limit of detection (LOD) has not been recognized. The terms ‘LOD' or 'limit of detection' cannot be spotted on either Figure 2a or Figure 2b. In fact, the term ‘LOD' does not appear in the tweets before retraction and only 19 times since retraction. The term ‘limit of detection' occurred once before retraction (on the evening before where it probably has been announced already) and seven times after retraction. Three tweets mentioned both, ‘LOD' and 'limit of detection'. In the case of Bae, et al. (2020a), we can conclude that the reason for retraction has not been mentioned by Twitter users before a retraction decision has been made and only very rarely (0.5%) after publication of the retraction.

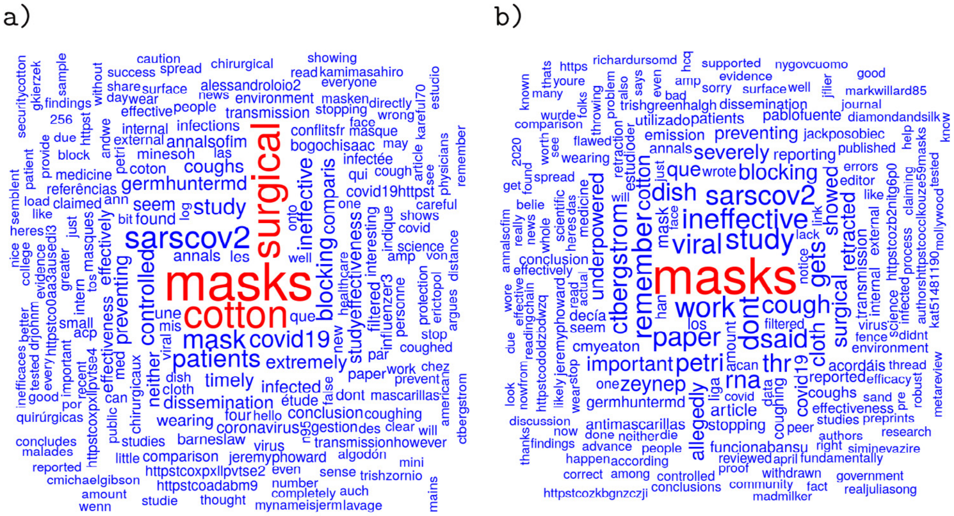


Figure 2: Word cloud from tweet texts based on the tweets mentioning Bae, et al. (2020a) before publication date of the retraction in panel a and based on the tweets mentioning Bae, et al. (2020a) or Bae, et al. (2020b) since publication date of the retraction in panel b

Twitter activity of Wang, et al. (2020a) and Wang, et al. (2020b)

Wang, et al. (2020a) reported that “SARS-CoV-2 infects T lymphocytes through its spike protein-mediated membrane fusion”. The peer review process was very fast: the paper has been submitted on 21 March 2020, and accepted three days later on 24 March 2020. The paper has been published on 7 April 2020, and retracted on 10 July 2020 (Wang, et al., 2020b) because “[a]fter the publication of this article, it came to the authors attention that in order to support the conclusions of the study, the authors should have used primary T cells instead of T-cell lines. In addition, there are concerns that the flow cytometry methodology applied here was flawed. These points resulted in the conclusions being considered invalid.” In this case, the retraction was made because of methodological errors that were not discovered during the peer review process. Figure 3 shows the tweets per day that mentioned either the paper or its retraction. The publication dates of the paper and the retraction notice are marked as gray vertical lines. Most tweets mentioning the publication occurred before retraction. Only very few tweets mentioned the retraction notice (n=53) or the paper after retraction (n=50). The tweet texts are analyzed using two different time periods: (i) before the publication date of the retraction and (ii) since the publication date of the retraction.

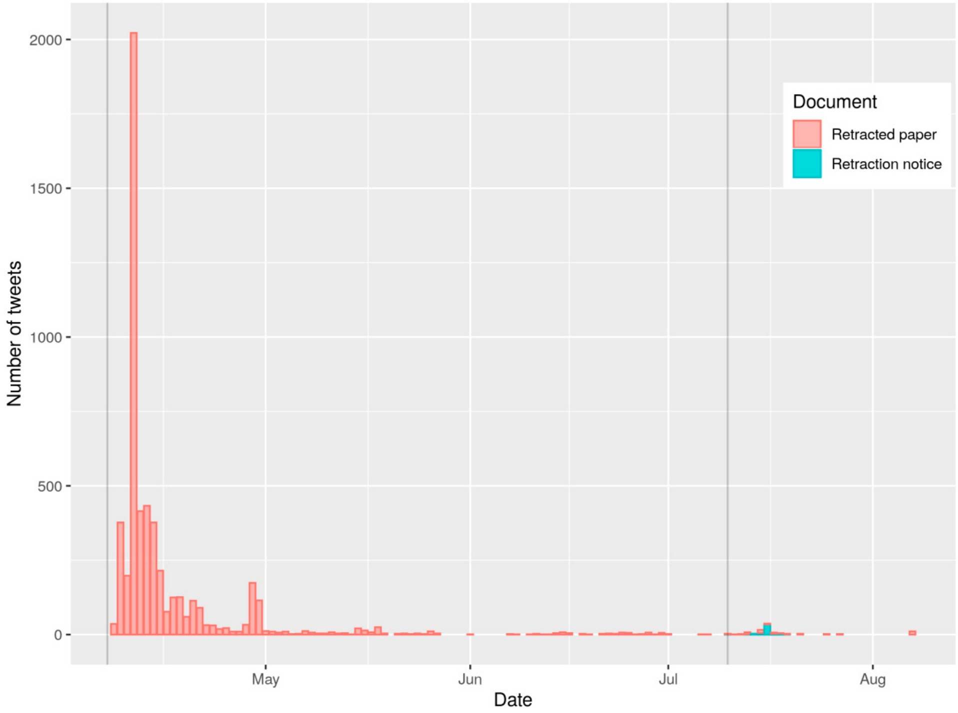


Figure 3: Tweets per day that mentioned either the paper Wang, et al. (2020a) or its retraction Wang, et al. (2020b). The publication dates of the paper and the retraction notice are marked as gray vertical lines.

Figure 4a shows a word cloud from tweet texts based on the tweets mentioning Wang, et al. (2020a) before the publication date of the retraction. Figure 4b shows a word cloud from tweet texts based on the tweets mentioning Wang, et al. (2020a) or Wang, et al. (2020b) since the publication date of the retraction. The most prominent terms in Figure 4a seem to be related to the virus and the disease it is causing whereas the term ‘retracted' is the most prominent one in Figure 4b.

One can easily see the terms ‘tcells’ in Figure 4a and Figure 4b, although not being among the most prominent terms. The study Wang, et al. (2020a) has been retracted because the authors should have used primary T cells instead of T-cell lines and because the flow cytometry methodology applied in the study was flawed. There were 5378 tweets registered before retraction and 103 since then. Only 14 tweets before retraction and four since then included the term ‘primary’ in the tweet text. The first tweet that mentioned problems with the used T cells was posted on 9 April 2020 (i.e., two days after publication and about three months before retraction of the study) and reads: “I would take this with a grain of salt. This was on cell lines, not primary T-cells. No evidence that this infection results in T-cell dysfunction. Furthermore, severe disease is characterized by hyper-immune response. Much is left to be desired here. Not sure what to conclude. https://t.co/15qvCa0fvw” (Sidholm, 2020).

There were 23 tweets that contained the term ‘flow cytometry’ and 101 tweets that contained the term ‘flowcytometry’ (mainly as a hashtag), both before publication of the retraction notice. Most of these 124 tweets (n=117) were retweets as identified by their first two letters ‘RT’ followed by a whitespace. The first tweet that casted doubt on the employed flow cytometry methodology was posted on 9 April 2020 and it reads: “Lots of discussion on Twitter about this paper, however some questions outstanding: Does the flow cytometry actually show cell infection? Where is the positive control comparison using classically infected cells?" (Vincent, 2020). Only a single tweet (retweeted 26 times) was found that mentioned 'flowcytometry' (as a hashtag) and none that mentioned 'flow cytometry' after publication date of the retraction notice.

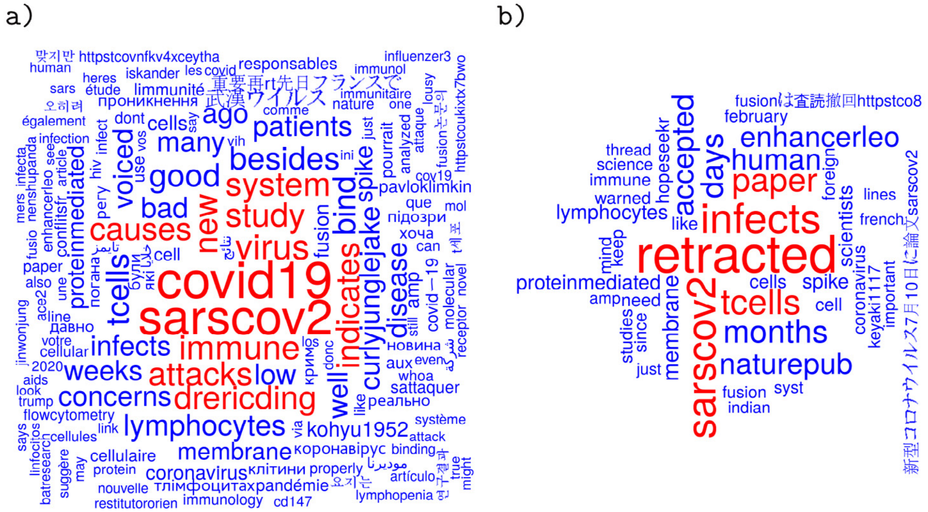


Figure 4: Word cloud from tweet texts based on the tweets mentioning Wang, et al. (2020a) before publication date of the retraction in panel a and based on the tweets mentioning Wang, et al. (2020a) or Wang, et al. (2020b) since publication date of the retraction in panel b

Twitter activity of Mehra, Desai, et al. (2020) and Mehra, Ruschitzka, and Patel (2020)

Probably the most attention among the three publications was drawn to the study by Mehra, Desai, et al. (2020). They reported that they could not confirm a benefit in COVID-19 treatment with hydroxychloroquine. They even reported that hydroxychloroquine increases the risk of complications during medical treatment against COVID-19. The study was published on 22 May 2020, and retracted on 05 June 2020 (Mehra, Ruschitzka, et al., 2020) because "several concerns were raised with respect to the veracity of the data and analyses conducted by Surgisphere Corporation and its founder" and co-author of the study. Surgisphere declined to transfer the full dataset to an independent third-party peer reviewer because that would violate client agreements and confidentiality requirements. Potential benefit or risk of hydroxychloroquine for treatment of COVID-19 is still not clear. In this case, the retraction was made because of doubts regarding the validity of the employed data that was not discovered in the peer review process.

Figure 5 shows the tweets per day that mention either the paper or its retraction. The publication dates of the paper and the retraction notice are marked as gray vertical lines. The retraction notice has been tweeted a day before its official publication date. Because the retraction notice has been tweeted a day before its official publication, the two time periods were organized slightly different: (i) before the day before the publication date of the retraction notice and (ii) since the day before the publication date of the retraction.

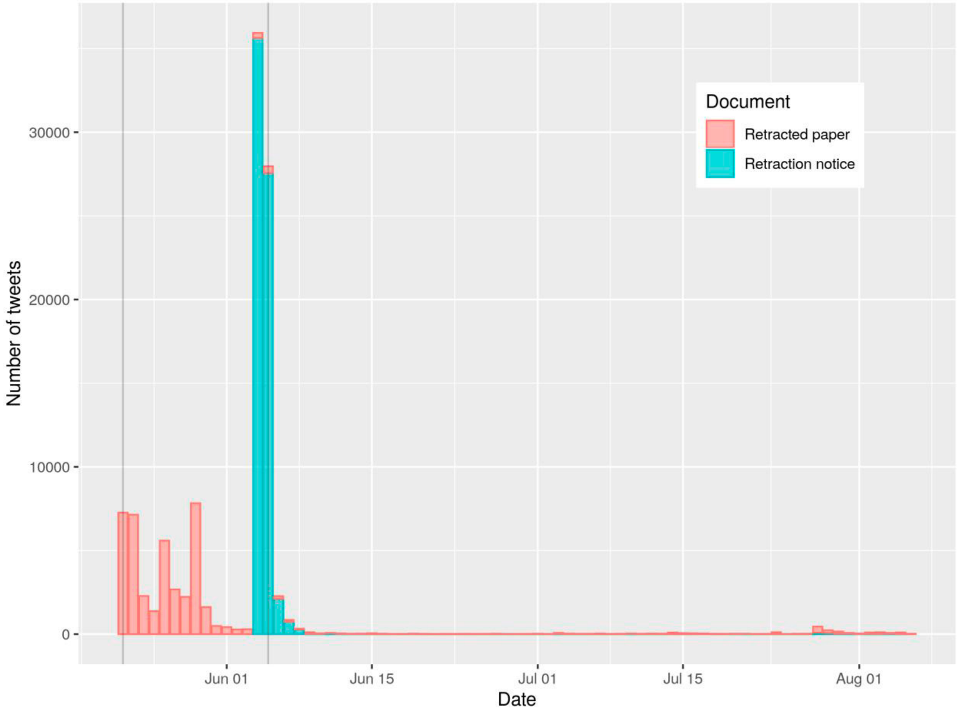


Figure 5: Tweets per day that mention either the paper Mehra, Desai, et al. (2020) or its retraction Mehra, Ruschitzka, et al. (2020). The publication dates of the paper and the retraction notice are marked as gray vertical lines.

Figure 6a shows a word cloud from tweet texts based on the tweets mentioning Mehra, Desai, et al. (2020) before the day before the publication date of the retraction. Figure 6b shows a word cloud from tweet texts based on the tweets mentioning Mehra, Desai, et al. (2020) or Mehra, Ruschitzka, et al. (2020) since the day before the publication date of the retraction. The most prominent terms in Figure 6a are 'lancet' (the journal the study was published in) and 'roaultdidier' (a French physician and microbiologist specializing in infectious diseases who promoted the hydroxychloroquine-based treatment of COVID-19). The most pronounced terms in Figure 6b are 'study', 'thelancet', and 'hydroxychloroquine'. The terms 'retraction' and 'retracted' are also quite prominent.

The publication Mehra, Desai, et al. (2020) was retracted because there was some doubt about the data basis from the company Surgisphere. The term 'surgisphere' appears at the lower border of Figure 6a. The term 'data' can be found in the right part of Figure 6b. The study Mehra, Desai, et al. (2020) was mentioned 39477 times on Twitter before the retraction notice was tweeted for the first time (4 June 2020). Out of these 39477 tweets, 720 contained the term 'surgisphere' and 1819 contained the term 'data'. The first ten tweets mentioning the study and using the term 'data' occurred on the day of publication of the study. None of them mentioned problems regarding the employed data. The first tweet mentioning the term 'surgishphere' was posted on the day of publication of the study (Stamets, 2020) but did not cast doubt on the validity of the data. A later tweet from 24 May 2020 (i.e., two days after publication and eleven days before retraction of the study) mentioning 'surgisphere' as a twitter handle (Arkancideisreal, 2020a) referenced a tweet that casted doubt on the validity of the data (Arkancideisreal, 2020b). Many later tweets mentioning the study and using the term ‘surgisphere’ posted between 29 May 2020 and 2 June 2020 also casted doubt on the employed data, see for example Pinjos (2020) and Schwartz (2020). Some of them linked to an open letter to the authors of the study and Richard Horton (Editor of *The Lancet*), see Watson (2020). Both terms, ‘data’ and ‘surgisphere’, occurred in 61 tweets.

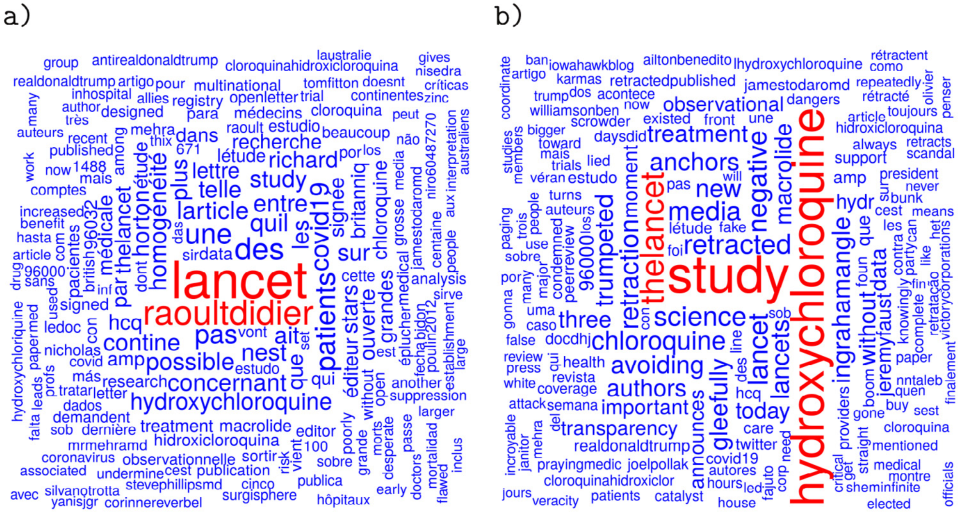


Figure 6: Word cloud from tweet texts based on the tweets mentioning Mehra, Desai, et al. (2020) before the day before the publication date of the retraction in panel a and based on the tweets mentioning Mehra, Desai, et al. (2020) or Mehra, Ruschitzka, et al. (2020) since the day before the publication date of the retraction in panel b

Discussion

Methodological mistakes, data errors, and scientific misconduct are prevalent problems in science that are often difficult to detect. In this case study, we are interested in the question whether tweets can be used to detect problems early with scientific papers. We have analyzed three different papers about COVID-19/SARS-CoV-2 that were retracted at least two weeks after publication.

Mixed conclusions can be drawn from this case study. Our results show that not all problems of publications can be spotted using Twitter data. However, one can find hints to problems regarding publications in tweet texts. In the tweets that mentioned Bae, et al. (2020a), we were not able to find early warnings regarding problems with the publication. Only a tweet on the night before the retraction occurred, was connected to the retraction reason. We were able to find early warning signs in tweet texts that mentioned the other two retracted publications of our study. However, the early warning signs from @John_Will_I_Am regarding the study by Wang, et al. (2020a) are less compelling than the clear warnings from @Arkanicide_is_real regarding the study by Mehra, Desai, et al. (2020).

It is noteworthy, that tweet texts have to be analyzed because many tweets mentioning a publication are not a clear indicator of problems with a publication. Performing more and larger case studies like ours might yield a list of Twitter users who often mention problems with publications early on. Many case studies might lead in a meta-analysis to a curated list of Twitter users who often spotted problems in publications in the past. Such a curated list of Twitter users might be helpful for revealing problems in future publications earlier. Maybe also more problematic publications can be found with such a Twitter-based discovery system than without such a system. Besides monitoring the activity of special Twitter users, a search term list of alarming keywords might be helpful. Such a list will be subject-related and differ from one field to another. Experts of the fields should propose or at least check such a list. Limiting such monitoring to publications that are also discussed on post-publication peer review sites, such as PubPeer, or in blogs might provide a more focused monitoring.

Acknowledgments

IDs of tweets mentioning the studied documents were retrieved from the Altmetric.com API. Tweets with their full text and posting dates were retrieved from the Twitter API. We thank Vanash Patel (Imperial College London) for checking our manuscript for potential misrepresentations of medical sciences. However, any problematic statements are the sole responsibility of the authors.

References

Arkancideisreal. (2020a). Tweet. Retrieved 25 August 2020, from https://twitter.com/Arkancideisreal/status/1264667066813935616

Arkancideisreal. (2020b). Tweet. Retrieved 25 August 2020, from https://twitter.com/Arkancideisreal/status/1264032084944814082?s=20

Bae, S., Kim, M.-C., Kim, J. Y., Cha, H.-H., Lim, J. S., Jung, J., . . . Kim, S.-H. (2020a). Effectiveness of surgical and cotton masks in blocking SARS–CoV-2: A controlled comparison in 4 patients. *Annals of Internal Medicine, 173*(1), W22-W23. doi: 10.7326/M20-1342.

Bae, S., Kim, M.-C., Kim, J. Y., Cha, H.-H., Lim, J. S., Jung, J., . . . Kim, S.-H. (2020b). Notice of retraction: Effectiveness of surgical and cotton masks in blocking SARS-CoV-2. *Annals of Internal Medicine, 173*(1), 79. doi: 10.7326/L20-0745.

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies, 1*(1), 363-376. doi: 10.1162/qss_a_00018.

Bouchet-Valat, M. (2014). SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library. R package version 0.5.1. Retrieved 11 August 2020, 2020, from https://CRAN.R-project.org/package=SnowballC

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences, 109*(42), 17028. doi: 10.1073/pnas.1212247109.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, 25*(5), 1-54.

Fellows, I. (2014). Wordcloud: Word clouds. R package version 2.5. Retrieved 11 August 2020, 2020, from https://CRAN.R-project.org/package=wordcloud

Haustein, S. (2019). Scholarly Twitter metrics. In W. Glänzel, H. F. Moed, U. Schmoch & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 729-760). Cham, Switzerland: Springer International Publishing.

Hornik, K. (2014). NLP: Natural language processing infrastructure. R package version 0.1-5. Retrieved 11 August 2020, 2020, from https://CRAN.R-project.org/package=NLP

Lang, D. T., & the CRAN team. (2018). RCurl: General network (HTTP/FTP/...) client interface for R, from https://CRAN.R-project.org/package=RCurl

Mehra, M. R., Desai, S. S., Ruschitzka, F., & Patel, A. N. (2020). Retracted: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: A multinational registry analysis. *The Lancet*. doi: 10.1016/S0140-6736(20)31180-6.

Mehra, M. R., Ruschitzka, F., & Patel, A. N. (2020). Retraction - hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: A multinational registry analysis. *The Lancet, 395*(10240), 1820. doi: 10.1016/S0140-6736(20)31324-6.

Müller, K., Wickham, H., James, D. A., & Falcon, S. (2017). RSQLite: 'SQLite' interface for R. R package version 2.0. Retrieved 22 June 2020, from https://CRAN.R-project.org/package=RSQLite

Neuwirth, E. (2014). RColorBrewer: ColorBrewer palettes. R package version 1.1-2. Retrieved 11 August 2020, 2020, from https://CRAN.R-project.org/package=RColorBrewer

Pinjos, L. (2020). Tweet. Retrieved 24 August 2020, from https://twitter.com/LPinjos/status/1266639081720754176

R Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

R Special Interest Group on Databases (R-SIG-DB), Wickham, H., & Müller, K. (2018). DBI: R database interface.

Schwartz, I. (2020). Twitter. Retrieved 24 August 2020, from https://twitter.com/GermHunterMD/status/1267496755702054917

Sidholm, J.-W. (2020). Tweet. Retrieved 24 August 2020, from https://twitter.com/John_Will_I_Am/status/1248328733397671936

Sotudeh, H., Barahmand, N., Yousefi, Z., & Yaghtin, M. (2020). How do academia and society react to erroneous or deceitful claims? The case of retracted articles' recognition. *Journal of Information Science*, 0165551520945853. doi: 10.1177/0165551520945853.

Stamets, B. (2020). Tweet. Retrieved 25 August 2020, from https://twitter.com/BillStamets/status/1263845285803016202

Sugawara, Y., Tanimoto, T., Miyagawa, S., Murakami, M., Tsuya, A., Tanaka, A., . . . Narimatsu, H. (2017). Scientific misconduct and social media: Role of twitter in the stimulus triggered acquisition of pluripotency cells scandal. *J Med Internet Res, 19*(2), e57. doi: 10.2196/jmir.6706.

Teixeira da Silva Jaime, A., & Dobránszki, J. (2019). A new dimension in publishing ethics: social media-based ethics-related accusations. *Journal of Information, Communication and Ethics in Society, 17*(3), 354-370. doi: 10.1108/JICES-05-2018-0051.

Toupin, R. (2020). twitterprofiles. Retrieved 22 June 2020, from https://github.com/toupinr/twitterprofiles

Toupin, R., & Haustein, S. (2018). *A climate of sharing: Who are the users engaging with climate research on Twitter*. Paper presented at the altmetrics18 Workshop at the 5:AM Conference, London, UK. https://doi.org/10.6084/m9.figshare.7166393.v1

Toupin, R., Millerand, F., & Larivière, V. (2019). *Scholarly communication or public communication of science? Assessing who engage with climate change research on Twitter.* Paper presented at the 17th International Conference on Scientometrics and Informetrics (ISSI 2019) with a special STI conference track, Rome, Italy.

Vainio, J., & Holmberg, K. (2017). Highly tweeted science articles: Who tweets them? An analysis of Twitter user profile descriptions. *Scientometrics, 112*(1), 345–366. doi: 10.1007/s11192-017-2368-0.

Vincent, B. (2020). Tweet. Retrieved 24 August 2020, from https://twitter.com/BenjaminGVincen/status/1248243908837965828

Wang, X., Xu, W., Hu, G., Xia, S., Sun, Z., Liu, Z., . . . Lu, L. (2020a). Retracted article: SARS-CoV-2 infects T lymphocytes through its spike protein-mediated membrane fusion. *Cellular & Molecular Immunology*. doi: 10.1038/s41423-020-0424-9.

Wang, X., Xu, W., Hu, G., Xia, S., Sun, Z., Liu, Z., . . . Lu, L. (2020b). Retraction Note to: SARS-CoV-2 infects T lymphocytes through its spike protein-mediated membrane fusion. *Cellular & Molecular Immunology, 17*(8), 894-894. doi: 10.1038/s41423-020-0498-4.

Watson, J. (2020). An open letter to Mehra et al and The Lancet. Retrieved 24 August 2020, from https://zenodo.org/record/3862789#.X0PgGc9CSUn

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag

Wickham, H. (2017a). Httr: Tools for working with URLs and HTTP, from https://CRAN.R-project.org/package=httr

Wickham, H. (2017b). Tidyverse: Easily install and load the 'Tidyverse'. R package version 1.2.1. Retrieved 22 June 2020, from https://CRAN.R-project.org/package=tidyverse

Zuckerman, H. (2020). Is "the time ripe" for quantitative research on misconduct in science? *Quantitative Science Studies, 1*(3), 945-958. doi: 10.1162/qss_a_00065.

Examining International Research Collaboration during the COVID-19 Pandemic using arXiv Preprints

Jiangen He[1], Erjia Yan[2] and Chaoqun Ni[3]

[1] *jiangen@utk.edu*
School of Information Sciences, The University of Tennessee, Knoxville (United States)

[2] *erjia.yan@drexel.edu*
College of Computing and Informatics, Drexel University, Philadelphia (United States)

[3] *chaoqun.ni@wisc.edu*
The Information School, University of Wisconsin-Madison, Madison (United States)

Abstract

This research-in-progress paper seeks to understand how the pandemic affected international research collaboration different countries and regions. It collected 333,793 preprints submitted to ArXiv between 2019 and 2020 to compare international research collaboration patterns pre-COVID-19 and COVID-19 eras. The paper finds that international research collaboration has been substantially affected by the pandemic, but the impact is manifested in varied extent over different time periods and in different countries. The project observed 1.55% decrease in international research collaboration in 2020 as compared with 2019. More specifically, there was a significant drop of international research collaboration at the early stage of the pandemic (from January 2020 until May 2020), and a sturdy recovery (to pre-COVID time) after May 2020. The change pattern varies by discipline and country. The results also demonstrate the resilience and adaptiveness of the scientific community in maintaining international research collaboration.

Introduction

International Research Collaboration (IRC) is a powerful driving force for innovation and discovery, as witnessed by the most rapid vaccine development in history: many existing COVID-19 vaccines are the result of IRC. IRC has also been found to increase research productivity (Thelwall & Maflahi, 2020), and impact (Wagner & Jonkers, 2017) for researchers. With the pandemic affecting almost every aspect of society, efforts have been invested in investigating its impact on research productivity (Vincent-Lamarre, Sugimoto, & Larivière, 2020), journal submission patterns (Maghfour, Olson & Jacob, 2020), and how scientists collaborated on COVID-19 specific research (Fry, Cai, Zhang & Wagner, 2020; Haghani & Bliemer, 2020). Yet, there's a lack of research examining the impact of the pandemic on IRC patterns. This study aims to fulfill the gap by examining the IRC pattern under the pandemic using submissions to preprint servers.

IRC is a powerful force in improving research productivity (Thelwall & Maflahi, 2020) and impact (Wagner & Jonkers, 2017). Many recent studies have investigated IRC in coronavirus research (Fry, Cai, Zhang & Wagner, 2020; Haghani & Bliemer, 2020) and the impact of the pandemic on research productivity (Vincent-Lamarre, Sugimoto & Larivière, 2020) and journal submission patterns (Maghfour, Olson, & Jacob, 2020). However, the impact of the pandemic on IRC is not clear. This research-in-progress study aims to fill the gap by examining the impact of COVID-19 on IRC using preprint publications.

We collected all preprints submitted to aXriv.org between 2019 and 2020. We used computational methods to extract author affiliations and countries from submitted PDF files. Using 2019 preprint data as the baseline, the study compared the level of IRC during the pandemic at the author-, country-, and discipline-levels to identify the impact of the pandemic on IRC. We aim to answer two research questions:

1. Whether and how IRC changed under the COVID-19 pandemic?
2. Whether and how the change (if any) varies by discipline and country?

Data Collection

Coauthorship is often used to approximate research collaboration. This study considers arXiv.org submissions with more than one author and one country for their affiliations as IRC. Using submissions from preprint servers like arXiv.org is advantageous: it allows us to analyze research output without having to wait for publication delay. Using preprint submissions from arXiv.org, this study leveraged the capabilities of machine learning and cloud computing to extract metadata and geographic information from PDF files. Figure 1 shows the process of data collection.

1. Download the complete set of processed arXiv PDF files for 2019 and 2020 through arXiv bulk data access that is available from Amazon S3[1].
2. Convert PDF files into XML files in NLM JATS format by using an adapted version of CERMINE, a Java library and a web service for extracting metadata and content from PDF files[2].
3. Extract author affiliation information (including country) from NLM JATS XML files.
4. Map country information into *ISO 3166-1 alpha-3* code (i.e., a set of standard three-letter country codes) using a predefined Python dictionary. If a country code cannot be identified, country, address, and affiliation information is used to request geographic coordinates through Google Geocoding API[3]. Country information is extracted from geographic coordinates.
5. Download arXiv metadata through OAI protocol for metadata harvesting.
6. Extract author and discipline information from arXiv metadata.

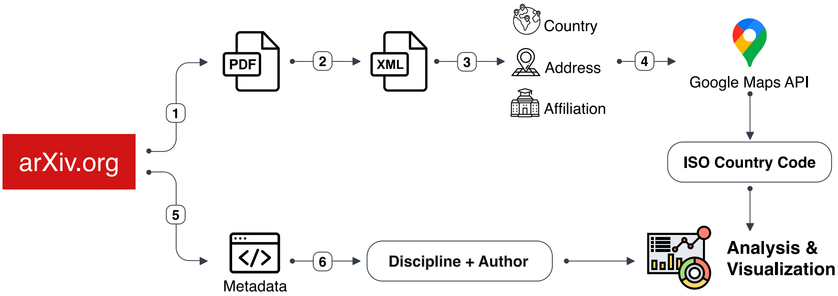


Figure 1. The process of data collection

We collected 155,464 and 178,329 preprints submitted to arXiv in 2019 and 2020, respectively. Country information for author affiliations was not available for some submissions due to missing, typo, etc. Figure 2 shows the number of collected preprints and the percentage of preprints with and without detected country information for each month. Country information was not detected for approximately 20% of harvested preprints. The percentage of preprints missing country information is roughly the same across months. Since most of our analysis is conducted by months, we assume limited impact from the missing on our conclusion. Ultimately, the final analytical sample include 240,648 preprints, 111,257 from 2019 and 128,391 from 2020.

The productivity patterns were also similar between the two years, but different in several months. June was the most productive month in 2020, but October was the one in 2019. The burst might be because scientific activities resumed and rebounded after the first wave of the pandemic in Spring 2020.

[1] https://arxiv.org/help/bulk_data_s3
[2] https://github.com/CeON/CERMINE
[3] https://developers.google.com/maps/documentation/geocoding/start

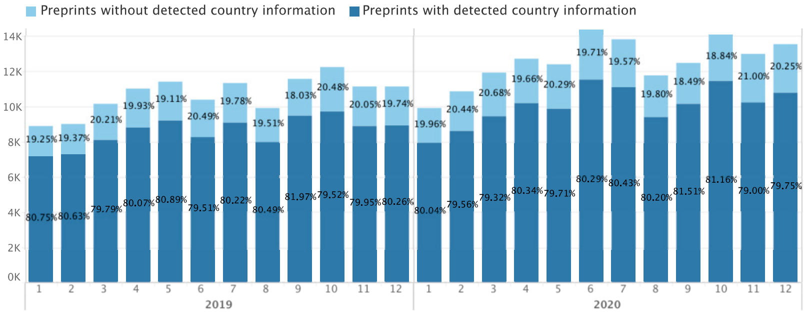


Figure 2. Collected data

Result

Collaboration at the author level

We first analyzed the collaboration at the author level. Figure 3 shows the percentage of papers with multiple authors. 90.34% (99,512) and 89.44% (115,990) of preprints had coauthors in 2019 and 2020, respectively. Compared with coauthorship patterns in 2019, more preprints in 2020 were authored by more than two, three, or four authors, and the pattern was consistent across 12 months. For example, 69.72% of preprints (89,513) were authored by at least three authors in 2020, but 67.95% of preprints (75,598) were authored by at least authors in 2019. We didn't see any evidence showing that multiple-author collaboration was affected by the pandemic.

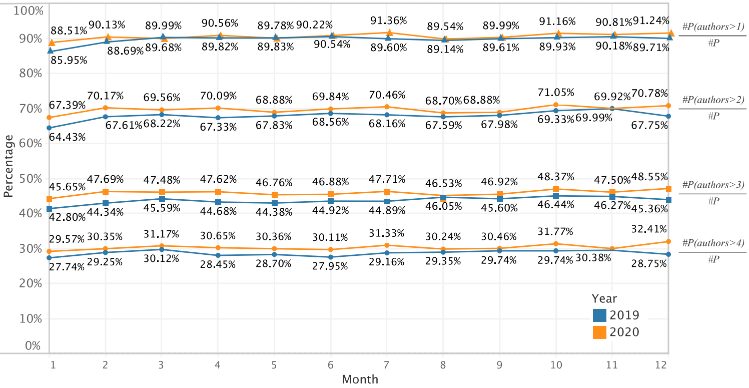


Figure 3. Collaboration at author level. #*P* is the number of papers, and #*P(author>n)* is the number of papers with more than *n* authors.

Collaboration at the country level

The IRC patterns were different between 2019 and 2020. In 2019, 38.81% of collaborative (more than one author) preprints (37,960) was contributed by authors from more than one country. This number declined to 37.26% (43,219) in 2020. Figure 4 shows the percentages of internationally collaborated preprints involving more than one, two, and three countries. Compared with 2019, we observed decreases in internationally collaborated preprints during the first four months of 2020. Since then, the percentage of IRC in 2020 was similar yet slightly lower to that of 2019. One possible explanation for this is that scientists may have adapted to new models of collaboration under the pandemic after the chaos and fluster in the first few months. Universities have been reported to invest more in online meeting tools and other capacities to ensure better communication and collaboration for both students and researchers. By comparing the results presented in Figure 3 and Figure 4, we found changes of IRC intensity in 2019 and 2020 (especially in the first few months of 2020, see Figure 4), though the changes were not noticeable at the author level when both domestic and IRCs are combined (Figure 3). The results suggest that when international collaborations were affected by the pandemic, scientists may have participated more in domestic collaborations to compensate for the impact of the pandemic on their research agenda.

The lower three panels in Figure 4 show the change of IRC with two, three, and more than three countries involved in 2019 and 2020. We observe that the most significant decline happened in the IRC involving more than three countries, especially for the first three months of 2020. This was likely due to the different levels of coordination needed based on the number of countries involved, and the different developing patterns of COVID-19 in different countries during that time.

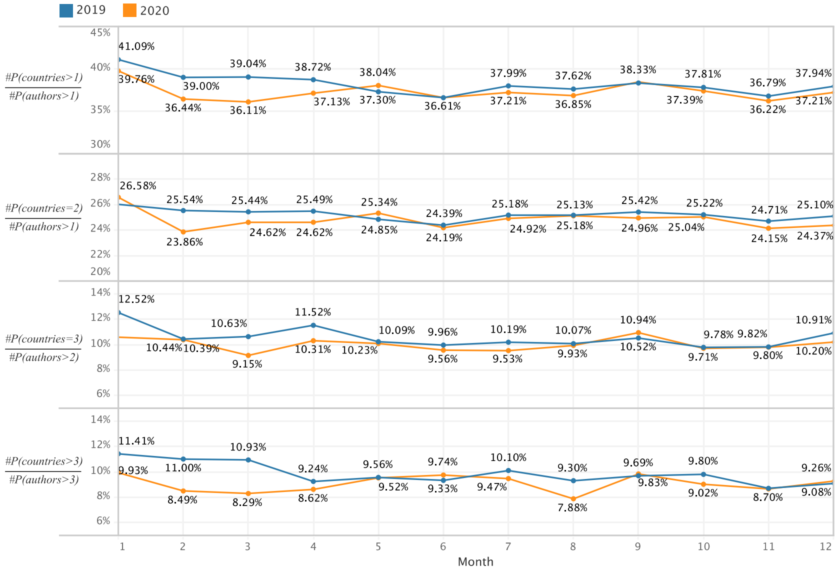


Figure 4. Collaboration at the country level. *#P(countries>n)* is the number of papers with authors from more than *n* countries.

Disciplines

We also compared the changes of IRC under the pandemic across disciplines. arXiv includes preprints from eight disciplines, with the majority of preprints (91.1%) are in physics (33.2%), computer science (29.3%) mathematics (20.6%%), and statistics (8.0%).

Figure 5 shows the percentage of IRC (preprints with authors from at least two countries) over all collaborative preprints in these four disciplines. The patterns in computer science, math, and physics were similar: IRC decreased for the first few months in 2020 and recovered to a similar level in 2019 afterward. The pattern in statistics was slightly different: after the decrease at the beginning of 2020, the IRC rebounded to a higher level than that in 2019.

The time and the extent of being influenced by the pandemic was different across disciplines due to the different nature of the disciplines. In comparison with other disciplines, IRC in computer science was influenced earlier. It takes varying time for scientists in a different discipline to adapt to the pandemic to recover IRC to the level before pandemic. It is also worth noting that some disciplines tended to have even higher level of IRC in the second half of 2020.

For example, scientists in statistics had much more IRC after July 2020 than in 2019. Another observation was that all disciplines except for physics had a dramatic increase in IRC in October 2020, but we need to further investigate this concerted increase.

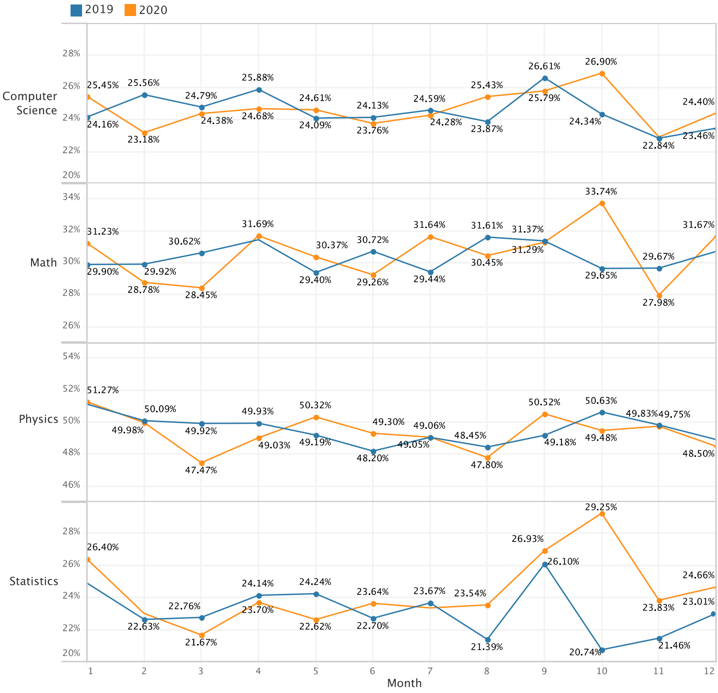


Figure 5. IRC in different disciplines.

Countries

Based on our analysis in previous sections, IRC was mainly hampered by the pandemic in the first four months of 2020. Thus, we examined the percentage change of IRC ($\#P(countries > 1)/\#P(authors > 1)$) in the first four months between 2019 and 2020 (see Figure 6). The IRC of most countries decreased, though with varied country-level differences. The IRC increased in some countries in Eastern Europe, Latin America, the Middle East, and Southeast Asia.

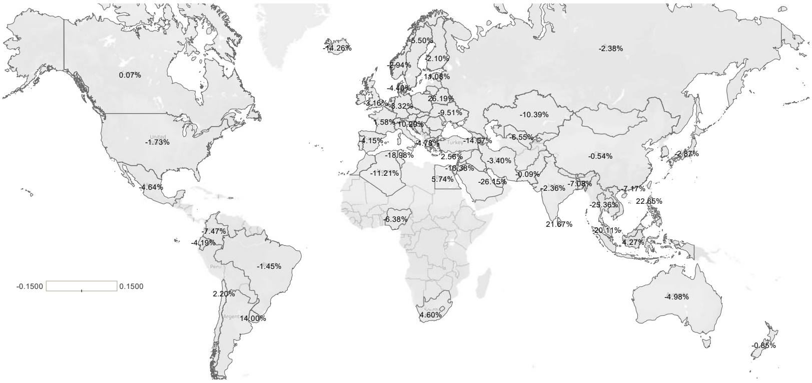


Figure 6. The change of percentage of IRC in the first four months between 2019 and 2020 across countries and regions submitted more than 100 arXiv preprints.

We also analyzed six countries with the most arXiv preprints (see Figure 7). The IRC decreased in the first few months in 2020 and recovered later. Some countries had a higher level of IRC in the second half of 2020 than 2019, such as the United Kingdom and Italy. The exception is France, where IRC was affected by the pandemic throughout the entire year in 2020.

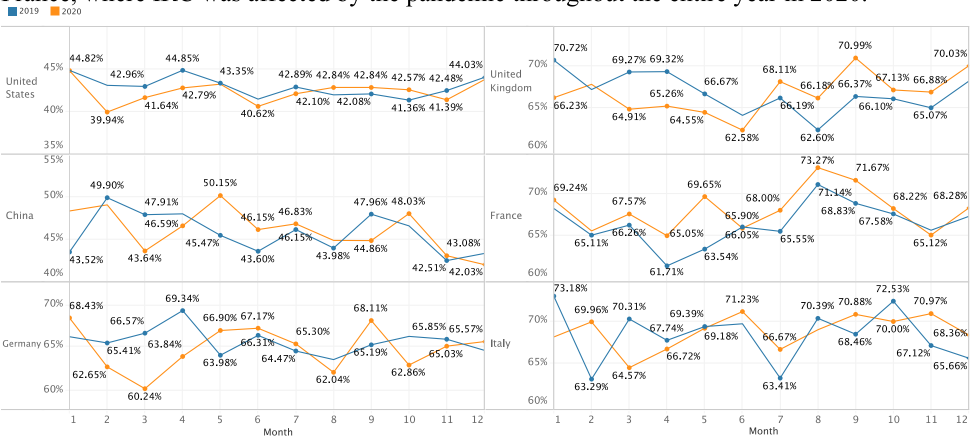


Figure 7. IRC from countries that submitted most preprints to arXiv.

Conclusion and Future Work

In this research-in-progress work, we found changes to IRCs in 2020, which is likely due to the COVID-19 pandemic. IRC rate declined during the first few months of 2020 and bounced back to a level similar to 2019 after May 2020. Yet, the change varied by discipline and country. Quick adoptions to research collaboration norms under the pandemic may have helped IRC resume during COVID-19. Different developing patterns as well as measures and efforts curtailing with COVID-19 by country is likely related to the varying changing patterns of IRC for countries.

Due to the varying impact of the pandemic across time, disciplines, and countries, we will examine factors affecting the impact of the pandemic in future research. For example, we will investigate the factors that alleviate or intensify the impact of the pandemic at the country level and examine the role of geographical and economic proximity in IRC during a public health crisis. This study, along with the future work, will provide a comprehensive and timely description of IRC patterns during the pandemic and provide an in-depth understanding of factors affecting IRC.

References

Fry, C. V., Cai, X., Zhang, Y. & Wagner, C. S. (2020). Consolidation in a crisis: Patterns of international collaboration in early COVID-19 research. *PLoS ONE*,15(7), e0236307.

Haghani, M. & Bliemer, M. C. J. (2020). Covid-19 Pandemic and the unprecedented mobilisation of scholarly efforts prompted by a health crisis: Scientometric comparisons across SARS, MERS and 2019-nCoV literature. *Scientometrics,* 125, 2695–2726.

Maghfour, J., Olson, J. & Jacob, S. E. (2020). COVID-19 impacts medical journal submissions. *International Journal of Women's Dermatology*, 6(4), 255–256.

Thelwall, M. & Maflahi, N. (2020). Academic collaboration rates and citation associations vary substantially between countries and fields. *Journal of the Association for Information Science and Technology*, 71(8), 968–978.

Vincent-Lamarre, P., Sugimoto, C. R. & Larivière, V. (2020). The decline of women's research production during the coronavirus pandemic. *Nature Index*, *News, 19 May.*

Wagner, C. S. & Jonkers, K. (2017). Open countries have strong science. *Nature*, 550(7674), 33.

National and Organizational Patterns of Nobel Laureate Careers in Physiology/Medicine, Physics and Chemistry.

Thomas Heinze[1] and Joel Emanuel Fuchs[2]

[1] *theinze@uni-wuppertal.de*
University of Wuppertal, Institute of Sociology & Interdisciplinary Center for Science and Technology Studies (IZWT), Gauss-Str 20, D-42119 Wuppertal (Germany)

[2] *jfuchs@uni-wuppertal.de*
University of Wuppertal, Institute of Sociology, Gauss-Str 20, D-42119 Wuppertal (Germany)

Abstract

This paper examines the distribution of Nobel laureates in Physiology/Medicine, Physics, and Chemistry across countries and research organizations. We provide basic information about where future laureates received their education and/or conducted their research, then present heatmaps depicting country and organizational specialization patterns. In addition, we identify the *organizational ultra-elite* in science: universities and research institutes that show continuously above-average numbers of future laureates, typically in one career phase. Furthermore, we identify those universities and research institutes that underwent considerable growth (or decline) in their capabilities for highly innovative research. Also, we compare country-specific profiles with those at the organizational level. Our findings are interpreted in the light of findings from comparative-historical studies.

Introduction

Following the seminal publication by Zuckerman (1977), the Nobel Prizes in Physiology/ Medicine, Physics, and Chemistry have attracted considerable attention from quantitative studies of science, especially with regard to achievement age (Jones & Weinberg, 2011; Redelmeier & Naylor, 2016), the time lag between prize-winning work and awarding of the prize (Becattini, Chatterjee, Fortunato, Pan, & Parolo, 2014; Fortunato, 2014), and the distribution of other science awards and collaboration networks in the years before and after their awarding (Chan, Gleeson, & Torgler, 2014; Chan, Önder, & Torgler, 2015). Several studies examined Nobel laureates from a bibliometric point of view, including spillover effects for the citations of laureates' publications unrelated to the Nobel Prize (Mazloumian, Eom, Helbing, Lozano, & Fortunato, 2011), and differences in citation patterns between landmark papers featuring theory, methods, experiments or inventions (Heinze, Heidler, Heiberger, & Riebling, 2013; Zhou, Xing, Liu, & Xing, 2014). Cross-national comparisons have found that Nobel Prizes "can be used to validate bibliometric indicators" (Rodríguez-Navarro, 2011).

More recently, analyses of the population of Nobel laureates have focused on the rise of North America as a global center of science and technology, its subsequent hegemony (Heinze, Pithan, & Jappe, 2019) and how national institutional contexts have shaped the capabilities of universities and research organizations to achieve scientific breakthroughs (Heinze, Heyden, & Pithan, 2020). These studies found that North America, in particular the United States, replaced Germany as global scientific center by the 1920s, that its hegemony was consolidated in the 1970s, and that although its leadership has come under pressure since the 2000s, a new global powerhouse is not in sight. Furthermore, it was shown that national contexts exerting weak institutional control are associated with organizational capabilities to achieve scientific breakthroughs. More specifically, countries with weak institutional control (United States, United Kingdom) have produced many more Nobel laureates, controlled by population size and by GDP per capita, than those exerting strong control (France, Germany).

However, much less attention has been paid to the distribution of laureates across universities and research institutes (Schlagberger, Bornmann, & Bauer, 2016). There is no comprehensive map of the organizational field in which future laureates were educated, conducted their prize-winning research, and worked when awarded the prestigious prize. Furthermore, rankings that include Nobel laureates, such as the "Academic Ranking of World Universities" (commonly known as Shanghai Ranking), do not consider where future Nobel laureates were educated or conducted their prize-winning research, but focus solely on information at the time when the Nobel Prize was awarded.

Data and Method

This paper examines the distribution of Nobel laureates in Physiology/Medicine, Physics, and Chemistry across national and organizational boundaries. We distinguish three career stages: (1) the university where future Nobel laureates received their highest academic degree (HD), (2) the university or research organization where they performed their prize-winning research (PWR), and (3) the university or research organization where they were employed at the time of the award (NP). Our analysis builds on an existing dataset (Heinze et al., 2020; Heinze, Pithan, et al., 2019) that has been updated and includes the entire time period 1901-2020 (120 years). The primary data source was the Nobel Foundation's website (www.nobelprize.org), enriched by data from secondary sources, such as the American Institute of Physics, American National Biography, Encyclopedia Britannica, Howard Hughes Medical Institute, National Academy of Sciences, Notable Names Database, and Royal Society.

First, we provide basic descriptive information about both the laureate population and the Top-50 universities and research organizations (Tab. 1). *Second*, we present heatmaps based on calculations of the specialization index RESP (see below). This index is calculated using the *Activity Index* (Narin, Carpenter, & Woolf, 1987; Piro et al., 2017), that captures the extent to which certain entities are specialized in certain activities (Formula 1). AI values lower 1.0 indicate a negative specialization (below-average scores), AI values greater 1.0 a positive specialization (above-average). A verbal expression of the AI, applied to Nobel laureates, is given in Formula 2.

Formula 1: General formula of the Activity Index (AI)

$$AI_{ij} := \frac{N_{ij}/\sum_i N_{ij}}{\sum_j N_{ij}/\sum_{ij} N_{ij}}$$

Formula 2: Specific AI applied to career phases of Nobel laureates

$$AI_{ij} := \frac{\text{Nobel laureates of University i in career phase j/Nobel laureates of University i}}{\text{All Nobel laureates in career phase j/All Nobel laureates}}$$

The AI's value range of $[0.0, +\infty]$ lacks an upper limit. Available indexes that are symmetrical both above and below the expected value include for example the *Revealed Symmetric Comparative Advantage (RSCA)* to capture country-specific technical specialization (Laursen, 2000, 2015). Furthermore, the *Relative Specialization Index (RSI*) has been used to map profiles of Scandinavian universities (Piro et al., 2011; 2017; 2014). Interpreting RSCA and RSI is easier than the AI due to their symmetrical value range of $[-1.0, +1.0]$: values lower 0.0 indicate negative specialization; values greater 0.0 indicate positive specialization.

We use a modified version of the RSCA/RSI index that was introduced by Grupp (1994, 1998). Its value range is $[-100.0, +100.0]$ with an expected value of zero (Formula 3). This index, which we call RESP (for "Index of Relative Specialization") is different from RSCA/ RSI in that it is based on the *hyperbolic tangent.* Consequently, its curve is steeper and reaches the upper limits of its value range more quickly than RSCA/RSI. Hence, RESP-based heatmaps are richer in contrast, and present more visibly specialization profiles. For further details, see Heinze, Tunger, Fuchs, Jappe, and Eberhardt (2019).

Formula 3: Relative Specialization (RESP)

$$\mathrm{RESP} \coloneqq 100\,\frac{\mathrm{AI}^2 - 1}{\mathrm{AI}^2 + 1}$$

Note: The subindices i and j of the AI are omitted for the sake of simplicity.

Results

Our dataset contains 1578 career events (HD, PWR, NP) and 341 organizations from 100 years (1901-2000), the latter including universities, public research institutes, and private research laboratories. The four countries with most career events (1229 or 78%) and most organizational entities (232 or 68%) are (in descending order): United States (729 career events & 120 orgs), United Kingdom (243 career events & 40 orgs), Germany (178 career events & 51 orgs), and France (79 career events & 21 orgs). We calculated RESP values based on all countries in the database (n=30), using 20-year periods. Figure 1 displays the results for the four countries.

Apart from the fact that there seems to be no clear career pattern for France, three results are noteworthy. *First*, the United States shows a decreasing specialization in educating future laureates (HD): compared to other countries, the United States increasingly relies on foreign-born and foreign-educated scientists. This result corroborates findings from Stephan and Levin (2001). At the same time, it becomes more specialized in the later career phases (PWR, NP), indicating its growing attractiveness over the 20th century as work environment for future laureates (Heinze et al., 2020). These developments are especially pronounced in the medical sciences. *Second*, Germany has almost the opposite specialization to that of the United States: it shows an increasing specialization in the education of future laureates (HD), whereas its attractiveness as work environment for later career phases has decreased in the second half of the 20th century. *Third*, the United Kingdom shows stability in the two later career phases: its specialization in PWR and NP is visible for the entire 20th century.

The first and second results are in line with comparative-historical evidence that highlights the declining hegemony of German universities in the early 20th century, coupled with an upswing of research universities in the United States (Ben-David, 1960, 1971). Ben-David explains this development both with regard to internal organizational features in North-American universities that were more conducive to the growth of new research fields (compared to those in Germany), and the more pronounced level of decentralized competition in the American university system, particularly between public and private universities (compared to exclusively public higher education in Germany). Also, the first and third results are in line with comparative-historical evidence suggesting that national contexts in United States and the United Kingdom exerted weak institutional control on universities and research organizations, and thus facilitated highly innovative research capabilities in the 20th century (Hollingsworth, 2004, 2006), a finding that is reflected also in data on institutional context in the 21st century (Pruvot & Estermann, 2017).

We turn now to the organizational level. Given the results above, it is certainly not astonishing that the Top-10 universities are from the United States (8) and the United Kingdom (2). Equally important, however, appears the considerable variation among the Top-50 with regard to their representation in the three career phases (Table 1). Therefore, we probed organizational specializations in Nobel laureates' careers. For this purpose, we calculated RESP values for all organizations in the database (n=341), using 20-year periods. Figures 2 and 3 display results for the Top-20. We also checked robustness by calculating RESP values for those organizations with more than two career events and for those with more than ten career events. Overall, specialization patterns were very robust. Therefore, we focus here on results for all organizations in the database. In our view, the following results are noteworthy.

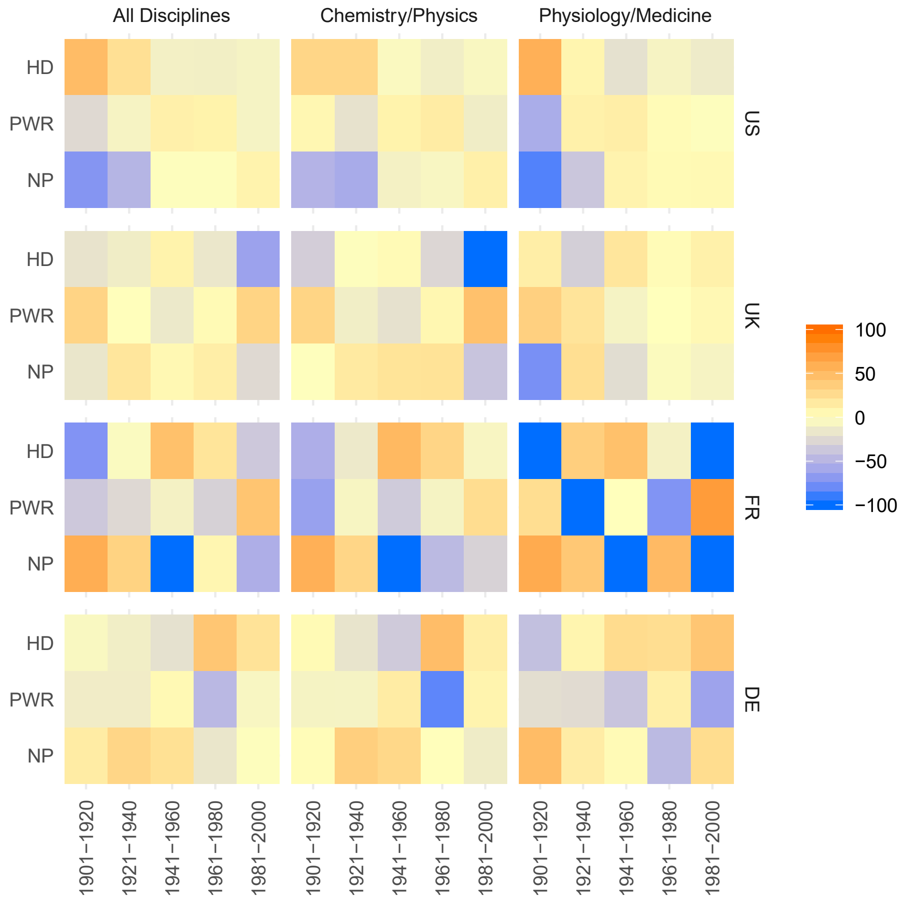


Figure 1. Career specialization profile of countries with Nobel laureates

First, there is some stability in single career phases over time, most notably in the education of future Nobel laureates (HD). Here, in at least four (out of five) consecutive periods, the following universities show a constant positive specialization over the 20th century: Cambridge, Harvard, Columbia, Berkeley, MIT, and Göttingen. Among those with a stable positive specialization in later career phases are in at least four (out of five) consecutive periods: Rockefeller (NP), Bell Labs (PWR), Caltech (NP), and London (NP). Clearly, constant positive specializations in either of PWR and/or NP require considerable resources to building and maintaining capabilities for highly innovative research. Borrowing a term coined by Zuckerman (1977), it is fair to call those universities and research institutes with constant positive specializations in either of the three career phases the *organizational ultra-elite* in global science. To be sure, this ultra-elite constitutes a very thin layer. Note that the Rockefeller Institute (later: Rockefeller University) stands out as the single entity with most 20-year periods in PWR and NP combined, highlighting its particular status among the organizational ultra-elite (for historical details on Rockefeller Institute see Hollingsworth, 2004).

Table 1. Global Top-50 universities and research organizations, 1901-2000

	Country	HD	PWR	NP	Total	Rank
University of Cambridge, Cambridge	UK	44	33	14	91	1
Harvard University, Cambridge	US	41	22	26	89	2
Columbia University, New York	US	21	19	9	49	3
University of California, Berkeley	US	20	15	10	45	4
California Institute of Technology, Pasadena	US	14	10	14	38	5
Massachusetts Institute of Technology, Cambridge	US	14	9	10	33	6
Rockefeller University, New York	US	1	14	16	31	7
University of Oxford, Oxford	UK	13	8	9	30	8
Stanford University, Palo Alto	US	6	9	14	29	9
Princeton University, Princeton	US	12	8	9	29	9
University of Chicago, Chicago	US	16	7	6	29	9
Cornell University, Ithaka	US	9	10	8	27	12
University of London, London	UK	6	7	12	25	13
Russian Academy of Sciences, Moscow	RU	6	8	9	23	14
Humboldt University, Berlin	DE	8	5	6	19	15
University of Goettingen, Goettingen	DE	11	3	5	19	15
Bell Laboratories, Murray Hill	US	0	15	4	19	15
University of Munich, Munich	DE	9	4	5	18	18
Washington University, St. Louis	US	3	11	4	18	18
MRC Laboratory of Molecular Biology, Cambridge	UK	2	7	7	16	20
University of Copenhagen, Copenhagen	DK	6	6	4	16	20
Karolinska Institute, Stockholm	SE	6	4	5	15	22
Yale University, New Haven	US	8	4	3	15	22
Johns Hopkins University, Baltimore	US	8	5	2	15	22
Swiss Federal Institute of Technology, Zurich	CH	5	5	3	13	25
Institut Pasteur, Paris	FR	1	5	6	12	26
Uppsala University, Uppsala	SE	3	4	5	12	26
University of Pennsylvania, Philadelphia	US	6	3	3	12	26
University Pierre and Marie Curie, Paris	FR	8	1	3	12	26
University of Illinois, Urbana-Champaign	US	6	4	2	12	26
Technical University, Munich	DE	5	4	2	11	31
University of Tokyo, Tokyo	JP	6	4	1	11	31
University of Heidelberg, Heidelberg	DE	3	1	6	10	33
University of Zurich, Zurich	CH	3	3	4	10	33
University of Wisconsin, Madison	US	5	2	3	10	33
Kyoto University, Kyoto	JP	3	5	2	10	33
University of Vienna, Vienna	AT	3	5	2	10	33
University of California, Los Angeles	US	2	3	4	9	38
Imperial College, London	UK	3	2	4	9	38
University of California, San Francisco	US	0	6	3	9	38
University of Washington, Seattle	US	2	4	3	9	38
University of Toronto, Toronto	CA	3	3	3	9	38
University of Strasbourg, Strasbourg	FR	4	3	2	9	38
University of Geneva, Geneva	CH	6	3	0	9	38
IBM Zurich Research Laboratory, Zurich	CH	0	4	4	8	45
National Institutes of Health (NIH), Bethesda	US	0	4	4	8	45
University of Texas Southwestern Medical Center, Dallas	US	1	3	4	8	45
University of Kiel, Kiel	DE	2	3	3	8	45
University of Freiburg, Freiburg	DE	2	4	2	8	45
University of Edinburgh, Edinburgh	UK	3	3	2	8	45
University of Rochester, Rochester	US	5	2	1	8	45
Nagoya University, Nagoya	JP	5	3	0	8	45

Note: HD=highest degree, PWR=prize-winning research, NP=award of Nobel Prize. Column "Total" sums up HD, PWR and NP.

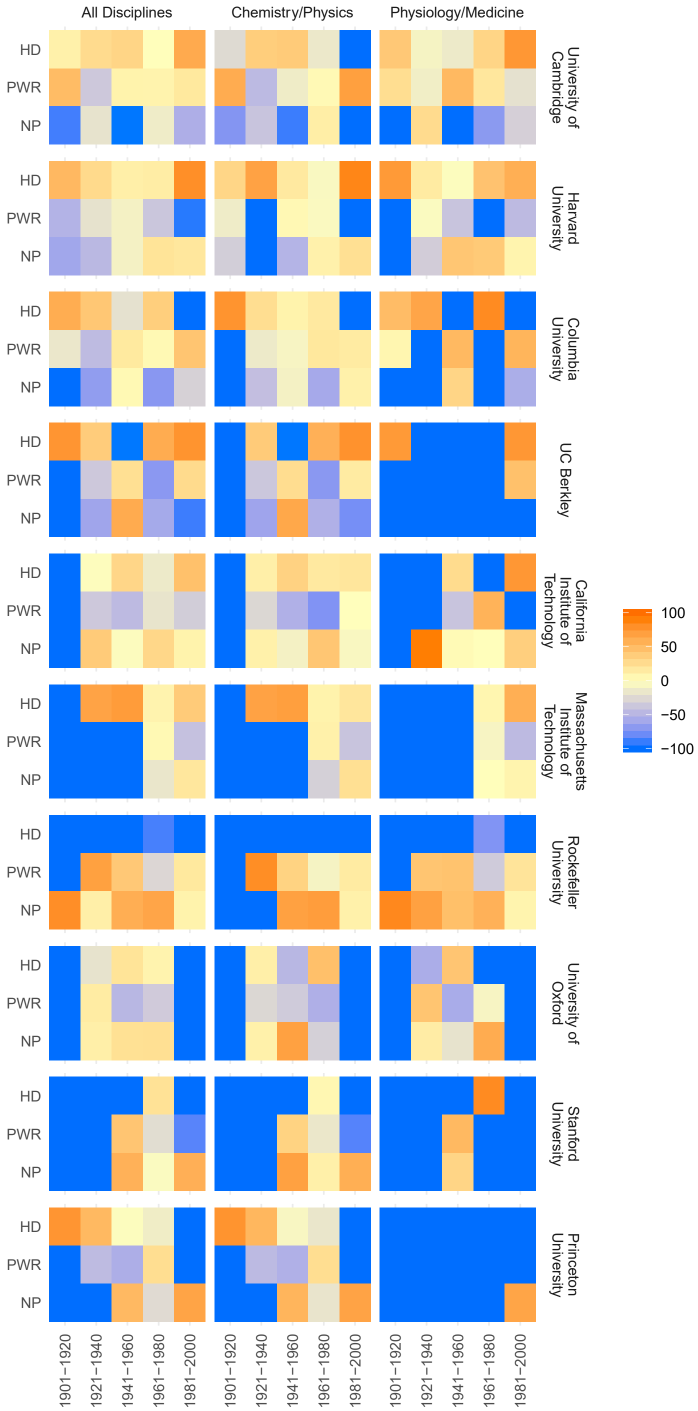


Figure 2. Nobel laureates career pattern of Top-10 universities and research organizations

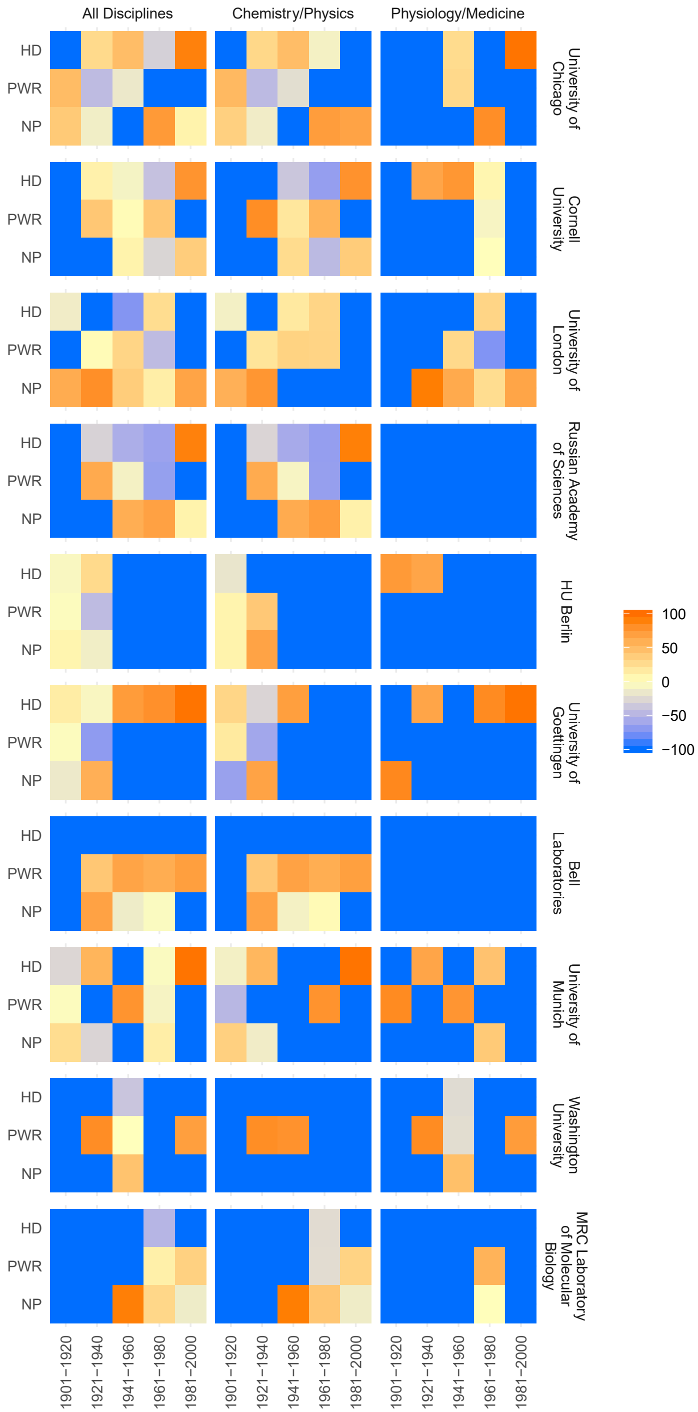


Figure 3. Nobel laureates career pattern of Top-11-20 universities and research organizations

Second, there is no single university or research organization with positive specializations in all three career phases in one (or more) 20-year period(s). More specifically, every university has at least one 20-year period, where no laureate either made his highest academic degree (HD), performed its price-winning research (PWR), or was employed at the time of the award (NP). However, we find examples that come close to that: MIT (1961-2000), Cambridge and Munich (1961-80), Columbia and Oxford (1941-60), HU Berlin and Göttingen (1901-20). Note that above-average scores in the PWR and NP career phases indicate capabilities for highly innovative research at an extremely high level. In addition, consider that some universities underwent a considerable change of their respective capabilities. For example, compare the first (1901-20) and fifth period (1981-2000) for two of the above-mentioned universities: HU Berlin (decrease) and MIT (increase). Also, most often several decades lie between HD and NP, so both HD and PWR observations are concentrated before 1980.

Third, few universities have changed their profile in a given career phase in one particular direction. Among those with growing specialization (over at least four consecutive periods) are Cambridge and Columbia (both PWR); conversely, among those with a decreasing specialization are Princeton (HD) and Göttingen (PWR). This suggests that in Cambridge and Columbia, some intra-organizational process of building-up capabilities to conduct highly innovative research took place, whereas in Princeton and Göttingen we assume that some process of downscaling of such capabilities occurred. How such processes unfolded and why is beyond the scope of this paper, but could be examined from a historical perspective.

In the light of the above-mentioned country-specific patterns, we probed whether they are reflected on the organizational level. Our analysis shows that this is not the case. *First*, there is no single university or research organization that roughly matches all three national specializations over time. Rather, we find some examples where specializations in one career phase (and sometimes two) are similar. Three examples to illustrate this point: 1) Cambridge mirrors the UK pattern in PWR (stable positive); 2) Princeton reflects the US pattern both in HD (decreasing) and NP (increasing), and 3) Göttingen develops a profile similar to that of Germany both in HD (increasing) and in NP (decreasing). *Second*, there are several universities that show patterns quite different from the national level. Two examples: 1) Cambridge is less specialized in the third career phase (NP) than the United Kingdom in general. Although it has educated an above-average number of future Nobel laureates and provided them with attractive working conditions, Cambridge retains them less often than the UK as a whole. 2) Similarly, although Columbia follows the (increasing) specialization of the United States in the first two career phases (HD, PWR), it has a weaker profile in NP compared to the national level.

Discussion

Taken together, our results suggest that analyzing longitudinal specialization patterns with regard to the careers of Nobel laureates yields several important insights complementing those obtained from cross-country comparisons. Interestingly, national career specialization patterns cannot be directly found on the organizational level. If universities and research institutes mirror national patterns, they do so in selected career phases only. However, the overall scientific growth (or decline) of countries can be seen in the profiles of particular universities and research institutes, as examples in Figures 2 and 3 illustrate for the United States and Germany. Perhaps the most important finding is the *organizational ultra-elite*, a group of (mostly private) universities and research institutes that show continuously above-average contributions in the education and employment of future laureates. In the light of commonly used rankings, such as the Shanghai Ranking or the Leiden Ranking that provide information about top-performing universities in the early 21st century, our analysis covers the entire 20th century, and thus gives insights into the building-up and maintenance of capabilities for highly innovative research.

In the future, our dataset can be analyzed further. Besides the RESP values for countries and for particular institutions, the interaction of both can be explored, i.e. how RESP values change in the context of world-wide versus country-wide consideration. Also, for most ultra-elite organizations, information about financial resources and scientific staff are available. The link between RESP values and both financial and human resources could expose more information about ultra-elite research organizations. Finally, we used descriptive statistic for presenting the RESP values here. Of course, RESP values can be statistically analyzed, too.

Notes

All graphs can be found and used under the CC-BY-NC-ND-4.0 international license at https://fachprofile.uni-wuppertal.de/conferences/issi2021.html. The Nobel Laureate dataset will be made publicly available following the completion of the ongoing research project.

Acknowledgments

We thank Marie von der Heyden for her support in data collection and data curation.

References

Becattini, F., Chatterjee, A., Fortunato, S., Pan, R. K., & Parolo, P. D. B. (2014). The Nobel Prize delay. *arXiv*, 1405.7136v1401.

Ben-David, J. (1960). Scientific Productivity and Academic Organization in Nineteenth Century Medicine. *American Sociological Review, 25*(6), 828-843.

Ben-David, J. (1971). *The Scientist's Role in Society. A Comparative Study*. Englewood Cliffs, N.J.: Prentice-Hall.

Chan, H. F., Gleeson, L., & Torgler, B. (2014). Awards before and after the Nobel Prize: A Matthew effect and/or a ticket to one's own funeral? *Research Evaluation, 23*, 210–220.

Chan, H. F., Önder, A. S., & Torgler, B. (2015). Do Nobel laureates change their patterns of collaboration following prize reception? *Scientometrics, 105*, 2215-2235.

Fortunato, S. (2014). Growing time lag threatens Nobels. *Nature, 508*, 186.

Grupp, H. (1994). The measurement of technical performance of innovations by technometrics and its impact on established technology indicators. *Research Policy, 23*, 175-193.

Grupp, H. (1998). Measurement with patent and bibliometric indicators. In H. Grupp (Ed.), *Foundations of the economics of innovation. Theory, Measurement, Practice* (pp. 141-188). Cheltenham: Edward Elgar.

Heinze, T., Heidler, R., Heiberger, H., & Riebling, J. (2013). New Patterns of Scientific Growth? How Research Expanded after the Invention of Scanning Tunneling Microscopy and the Discovery of Buckminsterfullerenes. *Journal of the American Society for Information Science and Technology, 64*, 829-843.

Heinze, T., Heyden, M. v. d., & Pithan, D. (2020). Institutional environments and breakthroughs in science. Comparison of France, Germany, the United Kingdom, and the United States. . *PLoS ONE, 15*(9), e0239805. .

Heinze, T., Pithan, D., & Jappe, A. (2019). From North American hegemony to global competition for scientific leadership. Insights from the Nobel population. *PLoS ONE, 14*(4), e0213916.

Heinze, T., Tunger, D., Fuchs, J. E., Jappe, A., & Eberhardt, P. (2019). *Research and teaching profiles of public universities in Germany. A mapping of selected fields*. Wuppertal: BUW.

Hollingsworth, J. R. (2004). Institutionalizing Excellence in Biomedical Research: The Case of The Rockefeller University. In D. H. Stapleton (Ed.), *Creating a Tradition of Biomedical Research. Contributions to the History of the Rockefeller University* (pp. 17-63). New York: Rockefeller University Press.

Hollingsworth, J. R. (2006). A Path-Dependent Perspective on Institutional and Organizational Factors Shaping Major Scientific Discoveries. In J. T. Hage & M. Meeus (Eds.), *Innovation, Science, and Institutional Change* (pp. 423-442). Oxford: Oxford University Press.

Jones, B. F., & Weinberg, B. A. (2011). Age dynamics in scientific creativity. *Proceedings of the National Academy of Sciences of the United States of America, 108*(47), 18910-18914.

Laursen, K. (2000). Do export and technological specialisation patterns co-evolve in terms of convergence or divergence? Evidence from 19 OECD countries, 1971–1991. *Journal of Evolutionary Economics, 10*, 415-436.

Laursen, K. (2015). Revealed comparative advantage and the alternatives as measures of international specialization. *Eurasian Business Review, 5*, 99-115.

Mazloumian, A., Eom, Y.-H., Helbing, D., Lozano, S., & Fortunato, S. (2011). How Citation Boosts Promote Scientific Paradigm Shifts and Nobel Prizes. *PLoS ONE, 6*, e18975.

Narin, F., Carpenter, M. P., & Woolf, P. (1987). Technological assessments based on patents and patent citations. In H. Grupp (Ed.), *Problems of measuring technological change* (pp. 107-119). Köln: TÜV Rheinland.

Piro, F. N., Aksnes, D. W., Knudsen Christensen, K., Finnbjörnsson, Þ., Fröberg, J., Gunnarsdottir, O., Sivertsen, G. (2011). *Comparing Research at Nordic Universities using Bibliometric Indicators. Policy Brief 4/2011*. Oslo: NordForsk.

Piro, F. N., Aldberg, H., Aksnes, D. W., Staffan, K., Leino, Y., Nuutinen, A., Sivertsen, G. (2017). *Comparing research at nordic higher education institutions using bibliometric indicators covering the years 1999-2014. Policy Paper 4/2017*. Oslo: NIFU.

Piro, F. N., Aldberg, H., Finnbjörnsson, Þ., Gunnarsdottir, O., Karlsson, S., Skytte Larsen, K., . . . Sivertsen, G. (2014). *Comparing Research at Nordic Universities using Bibliometric Indicators – Second report, covering the years 2000-2012. Policy Paper 2/2014*. Oslo: NordForsk.

Pruvot, E. B., & Estermann, T. (2017). *University Autonomy in Europe III. The Scorecard 2017*. Retrieved from Brussels:

Redelmeier, R. J., & Naylor, C. D. (2016). Changes in Characteristics and Time to Recognition of Medical Scientists Awarded a Nobel Prize. *JAMA, 316*(19), 2043.

Rodríguez-Navarro, A. (2011). Measuring research excellence. Number of Nobel Prize achievements versus conventional bibliometric indicators. *Journal of Documentation, 67*, 582-600.

Schlagberger, E. M., Bornmann, L., & Bauer, J. (2016). At what institutions did Nobel laureates do their prizewinning work? An analysis of biographical information on Nobel laureates from 1994 to 2014. *Scientometrics, 109*, 723-767.

Stephan, P. E., & Levin, S. G. (2001). Exceptional Contributions to US Science by the Foreign-Born and Foreign-Educated. *Population Research and Policy Review, 20*(1-2), 59-79.

Zhou, Z., Xing, R., Liu, J., & Xing, F. (2014). Landmark papers written by the Nobelists in physics from 1901 to 2012: a bibliometric analysis of their citations and journals. *Scientometrics, 100*, 329-338.

Zuckerman, H. (1977). *Scientific Elite: Nobel Laureates in the United States*. New York: Free Press.

Opening the black box of bibliometric mapping algorithms: Which algorithm should I choose?

Matthias Held

matthias.held@tu-berlin.de
Social Studies of Science and Technology, TU Berlin, Hardenbergstr. 16-18, 10623 Berlin, Germany

Abstract

The task of bibliometric mapping involves the usage of a data model together with an algorithm to analyze the data. In the last years, we learned more about the behavior of data models like co-citation or bibliographic coupling and their usefulness for certain tasks. From these studies, however, we learn relatively little about the impact of the choice of the algorithm. Usually, bibliometric mapping algorithms are imported from other disciplines, which means that they have not been developed for this particular purpose. We provide here a framework to analyze if the different assumptions on the data that these algorithms make correspond with the properties of topics they are trying to reconstruct.

Introduction

Ever since the advent of larger and larger networks which can be created from bibliometric data, researchers were confronted with the task to find patterns in these data with the help of algorithms. The reconstruction of scientists' topics is one of these tasks with some decades of history, and predicated upon the idea that these topics can somehow be reconstructed from the traces they leave in the scientists' publications' metadata, if only one chooses a suitable approach and data model.

For the choice of the data model, we have comparative studies which analyze their usefulness to reconstruct the phenomena in question (Shibata et al., 2009; Klavans & Boyack, 2017). What these studies share is that the data model and the algorithm are considered together. The choice of the algorithm, however, is usually not subject to discussion. While Šubelj et al. (2016) systematically do compare a large set of commonly used algorithms and apply them to bibliometric networks, we do not learn why the results differ and which algorithm is suitable for which bibliometric application. The major reasons for the common practice of using algorithms as black box could be: (1) There is tacit knowledge present about the algorithms nobody writes about. (2) We do not know which community detection algorithm to use since we do not know what we are searching for. (3) We are confronted with large networks and therefore simply choose the high-performance algorithms. (4) Most algorithms used come from other disciplines and can be used ready-made as black box. The abovementioned study compares the output of algorithms according to their statistical properties, but these analyses are decoupled from a real bibliometric application. Hence, we still lack studies that assess algorithms under bibliometric laboratory conditions. We need deeper analyses into the properties of the original purposes and assumptions of each algorithm (Gläser et al., 2017: 993-4), and assess their correspondence with the properties of topic structures we are trying to reconstruct.

We present here a framework to analyze properties of algorithms to assess their correspondence with the properties of topics which we derive from our theoretical definition. For applying the framework we chose four algorithms which all have been used in the bibliometric context before and been 'successfully' applied, including studies applying the Louvain algorithm (Glänzel & Thijs, 2017), the Leiden Algorithm (Colavizza et al., 2020), Infomap (Velden et al., 2017) and OSLOM (Šubelj et al., 2016).

First, we provide the derived topic properties along with the criteria to analyse the algorithms which we consider relevant for the task. Then, we provide first results of the analysis for the first algorithm and describe our planned future work.

Properties of topics

For science studies to be able to incorporate the results obtained from bibliometric topic reconstruction studies, both need a (common) theoretically derived definition of topics. Thus, we start with a definition.

We consider topics to emerge from coinciding interpretations and uses of some scientific knowledge by researchers. Havemann et al. (2017:1091) have defined topics as *"a focus on theoretical, methodological or empirical knowledge that is shared by a number of researchers and thereby provides these researchers with a joint frame of reference for the formulation of problems, the selection of methods or objects, the organisation of empirical data, or the interpretation of data"*.

From this definition, several properties of topics can be derived, which in turn shape the demands to approaches aiming to reconstruct topics (Table 1):

1. "Topics are local in the sense that they are primarily topics to the researchers whose decisions are influenced by them and who contribute to them. Methods for topic identification using only local information reconstruct this insider perspective, while methods using also global information construct a compromise between insider and outsider perspectives on topics" (ibid.). Thus, when the purpose of the analysis is the reconstruction of the perspective of the producers of the topics, an approach using mainly local information is to be used.

2. They can have various sizes, from a few researchers working on them, up to many, many more. Hence, methods to reconstruct these should not force a fixed size distribution.

3. One or more researchers can work on several topics, yet share the same method, theory or empirical object. Thus, topics do overlap, and hence reconstruction approaches have to allow the overlap of topics, including pervasive overlap and the detection of hierarchies in the topics.

4. Topics as shared frames of reference are characterized by cohesion, and separation from other topics occurs as secondary result from that. Hence, cohesion should be considered by algorithms. Because topics as shared frames of reference can connect knowledge in many different unique ways (Held et al., 2021), it follows also that the topics can take various structural forms.

Table 1: Topic properties and relevant criteria of bibliometric mapping algorithms.

Properties of topics	Criteria for algorithm analysis
Local	Use of local information
Size variation	Flexible size distribution
Overlapping	Finding overlaps and hierarchies
Cohesive / structurally diverse	Community definition, cohesion-separation trade-off
-	User's degrees of freedom

Usually the users of algorithms are provided some set screws to change the behaviour of the algorithm. The major set screws, which we sum up under the criterium *user's degrees of freedom,* are briefly analysed, since a deeper investigation would comprise a separate study. To our knowledge, these properties of topics and derived criteria for the analysis of algorithms have never been considered and applied before in the context of mapping.

Properties of algorithms

One frequently applied set of approaches in our field consists of optimization algorithms to find communities in bibliometric networks, where the found communities are often directly interpreted as topics[i] (Šubelj et al., 2016; Sjögårde & Ahlgren, 2018). Note that the choice of optimization algorithms is predicated on the idea – next to the assumption that there are topic structures in the networks at all - that these structures (however different they may be) can be detected with one single function that gets optimized and that these structures are best represented by communities (not all relevant larger structures in networks need to be communities, Newman (2012)) .

We take these assumptions as given in this study and, thus, now look at the above-mentioned criteria of several community detection algorithms in order be able to evaluate if these agree with the properties of topics.

Definition of community / separation and cohesion: Each of the tested algorithms is an optimization algorithm, which means that each one optimizes a predefined function, which in turn means that a certain idea of how a community 'looks like' is build (implicitly) into the algorithm and also into the optimization function. For communities to be optimal, they would comprise maximum cohesive, well separated components (in the extreme this means components of cliques) (Havemann et al., 2019). Because these structures are rarely found in networks, each community detection algorithm must find a compromise between separation of the communities and internal cohesion of same. This compromise is tightly linked to the definition of a community. The two problems belong together: *"For a proper community definition, one should take into account both the internal cohesion of the candidate subgraph and its separation from the rest of the network"* (Fortunato & Hric, 2016:6). And each available algorithm represents a different approach to this question. At this point a rather tricky situation seems to appear. Even though so many different algorithms can be found in the community detection literature, some of which have been used for bibliometric purposes, it seems unclear from the literature how to best solve the problem of cohesion and separation (irrespective of our topic reconstruction problem) in line with an agreed-upon definition of community. Different conceptions of how communities should "look like" are built into the algorithms. What all four algorithms have in common is that they detect assortative structures (Newman & Clauset, 2016), i.e. structures that form dense regions, which, at this point, we have to assume to be characteristic for topic structures. However, the question as to how a high cohesion in network communities could be identified has various answers (Havemann et al., 2019), including internal conductance (communities should be hard to split), high density (many links between nodes), high clustering coefficient (number of links in nodes' neighborhoods divided by possible links) (Yang & Leskovec, 2014) or a high value of the second eigenvalue of the community's Laplacian matrix (Tibély, 2012). Again, there is no agreed-upon answer to this question. This also becomes clear when we look at the common efficient algorithms used, like the four analyzed in here, that mainly focus on separation[ii], and the discussion on what cohesion would mean is generally hardly taken up in the literature (ibid.). What separation means, on the other hand, seems more agreed upon.

Most of these considerations usually assume non-overlapping communities. When communities are allowed to overlap (a lot), or are assumed to comprise core-periphery structures, then finding the best trade-off between separation and cohesion becomes a completely different question (Havemann et al., 2019:2).

Use of local information: We consider algorithms which consider the statistics of the entire network and optimize the result using mainly these global statistics as global algorithms. In this paper these would comprise Louvain and Leiden algorithm, as well as Infomap. Those algorithms which consider local statistics by exploring and expanding the local network structures (as introduced by Baumes et al. (2005)) we consider as local algorithms. In this paper OSLOM is an example of this group.

Finding overlaps and hierarchies: Does the algorithm construct overlapping communities, or can it be used to obtain these? Are hierarchies in the network detected or how can the algorithm be used to obtain these?

Flexible size distribution: Does the algorithm force a certain community size distribution? Or tend to create many small or many large clusters?

Users' degrees of freedom: Which decisions does the algorithm request from the user and to which degree does this contribute to the constructedness of the result? More degrees of freedom can be helpful if they support tuning an algorithm to the bibliometric task, while fewer degrees of freedom may force a certain result and prevent experimentation. At the same time many degrees of freedom may make the link between settings screws and the outcomes nontransparent.

Analysis of Algorithms

In Table 2 the first results of the algorithms' properties are provided.

Table 2: Brief overview of four algorithms' properties relevant for topic reconstruction.

	Community Definition / Separation-Cohesion	Local / Global	Overlapping / Hierarchies	Preference of Size	Users' Freedom
OSLOM	Statistically significant community, focus on separation	Local	Pervasive / yes, automatically	unknown	p-value
Louvain	Dense community, focus on separation	Global	No / possible to extract simple hierarchy	unknown	Resolution parameter
Leiden	Dense community, focus on separation	Global	No / possible to extract simple hierarchy	unknown	Resolution parameter
Infomap	Nodes' closeness, focus on separation	Global	Marginal	unknown	Several parameters

Order Statistics Local Optimization Method (OSLOM)

This algorithm has been developed by Lancichinetti et al. (2011), who are authors very active in network science. It is based on the idea to optimize a function that assesses the statistical significance of each cluster. OSLOM starts with a random selection of single nodes as seeds for clusters, and repeatedly adds significant neighbors to node to let the cluster grow. Significance is determined here via a comparison to a distribution that is based on a null model, i.e. a network without community structure. This is done repeatedly, until similar clusters are found again and again, then the algorithm stops. This way, overlapping clusters are found.

Definition of community / separation and cohesion
The test for statistical significance of each community is based on separation. Though OSLOM checks for internal cluster structure in a community (Lancichinetti et al., 2011:5), and decides for splitting or merging of communities, it cannot ensure internal cohesion.

Use of local information
The OSLOM algorithm can mostly be considered a local approach, but also uses global statistics. When OSLOM considers the significance of a cluster, it first considers the surrounding of each cluster to add significant nodes, which is one aspect it can be considered a local approach. However, the assessment of the environment of a cluster is done by comparing it to a global random null model (Lancichinetti et al., 2011: 2), viz. the configuration model, which is also used in many variants of modularity. So, the global statistics play an important role in this step. When each cluster is then labelled "significant" with respect to its surrounding, the inside of the cluster is assessed for significance, using no more the global model, but by comparing it to a null model within this cluster, which makes it a local assessment of cluster statistics. Another local property of OSLOM is its search for possible cluster merging candidates, i.e. clusters which appear to be very similar (which, according to the different iterations for the determination of significance, often overlap a lot), where it is assessed if the overlapping zone is another separate cluster, if everything is merged or if only the overlapping zone remains significant and is kept (Lancichinetti et al., 2011: 5).

Finding overlaps and hierarchies
As OSLOM goes iteratively over each cluster and might add "significant nodes", overlaps between the clusters can occur because a node might be considered significant from the perspective of more than one cluster. Pervasive overlaps can occur.
In principle, OSLOM is able to detect a poly-hierarchy in a network. It searches for the smallest clusters which can be considered significant, and then builds a new network from these clusters, where again each node is assessed for significance. Thus, it is continued until no more significant clusters are found in the next coarser level, and potentially several levels of hierarchy are then found (Lancichinetti et al., 2011: 5).

Flexible size distribution
OSLOM's preference for cluster sizes is still under investigation.

Degrees of freedom as user
One major decision that has to be made is about the significance level P. It has influence on the size of the communities found, lower values lead to larger communities (fewer), higher values to smaller ones (more communities).

Future Work
First, our theoretical analysis of the four algorithms according to our derived criteria will be completed. From this, however, there is still a lot to do in order to make robust suggestions as to the suitability of an algorithm for a certain bibliometric task. It has been shown that no algorithm is suitable to cluster all kinds of networks (Aldecoa & Marín, 2013). Thus, as second step, we intend to use bibliometric benchmark datasets on which we apply the algorithms; and conduct a deeper assessment of the structures these algorithms detect.

References

Aldecoa, R., & Marín, I. (2013). Exploring the limits of community detection strategies in complex networks. *Scientific Reports*, *3*(1), 2216. https://doi.org/10.1038/srep02216

Baumes, J., Goldberg, M., Krishnamoorthy, M., Magdon-Ismail, M., & Preston, N. (2005). Finding communities by clustering a graph into overlapping subgraphs. *IADIS AC*, *5*, 97–104.

Colavizza, G., Costas, R., Traag, V. A., van Eck, N. J., van Leeuwen, T., & Waltman, L. (2020). *A scientometric overview of CORD-19* [Preprint]. Scientific Communication and Education. https://doi.org/10.1101/2020.04.20.046144

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, *659*, 1–44. https://doi.org/10.1016/j.physrep.2016.09.002

Glänzel, W., & Thijs, B. (2017). Using hybrid methods and 'core documents' for the representation of clusters and topics: The astronomy dataset. *Scientometrics*, *111*(2), 1071–1087. https://doi.org/10.1007/s11192-017-2301-6

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, *111*(2), 981–998. https://doi.org/10.1007/s11192-017-2296-z

Havemann, F., Gläser, J., & Heinz, M. (2017). Memetic search for overlapping topics based on a local evaluation of link communities. *Scientometrics*, *111*(2), 1089–1118. https://doi.org/10.1007/s11192-017-2302-5

Havemann, F., Gläser, J., & Heinz, M. (2019). Communities as Well Separated Subgraphs With Cohesive Cores: Identification of Core-Periphery Structures in Link Communities. *ArXiv:1810.10497 [Physics]*, *812*, 219–230. https://doi.org/10.1007/978-3-030-05411-3_18

Held, M., Laudel, G., & Gläser, J. (2021). Challenges to the validity of topic reconstruction. *Scientometrics*. https://doi.org/10.1007/s11192-021-03920-3

Klavans, R., & Boyack, K. W. (2017). Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984–998. https://doi.org/10.1002/asi.23734

Lancichinetti, A., Radicchi, F., Ramasco, J. J., & Fortunato, S. (2011). Finding Statistically Significant Communities in Networks. *PLoS ONE*, *6*(4), e18961. https://doi.org/10.1371/journal.pone.0018961

Newman, M. E. J. (2012). Communities, modules and large-scale structure in networks. *Nature Physics*, *8*(1), 25–31. https://doi.org/10.1038/nphys2162

Newman, M. E. J., & Clauset, A. (2016). Structure and inference in annotated networks. *Nature Communications*, *7*(1), 11863. https://doi.org/10.1038/ncomms11863

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for Information Science and Technology*, *60*(3), 571–580. https://doi.org/10.1002/asi.20994

Sjögårde, P., & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics*, *12*(1), 133–152. https://doi.org/10.1016/j.joi.2017.12.006

Šubelj, L., van Eck, N. J., & Waltman, L. (2016). Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods. *PLOS ONE*, *11*(4), e0154404.

Tibély, G. (2012). Criterions for locally dense subgraphs. *Physica A: Statistical Mechanics and Its Applications*, *391*(4), 1831–1847. https://doi.org/10.1016/j.physa.2011.09.040

Velden, T., Yan, S., & Lagoze, C. (2017). Mapping the cognitive structure of astrophysics by infomap clustering of the citation network and topic affinity analysis. *Scientometrics*, *111*(2), 1033–1051. https://doi.org/10.1007/s11192-017-2299-9

Yang, J., & Leskovec, J. (2014). Structure and Overlaps of Ground-Truth Communities in Networks. *ACM Transactions on Intelligent Systems and Technology*, *5*(2), 1–35.

[i] Some researchers, however, advocate a strict differentiation between community detection and classification into topics.

[ii] The more recent Leiden algorithm corrects the Louvain algorithms' sole focus on separation by improving communities' cohesion in the process.

A framework for measuring the knowledge diffusion impact of interdisciplinary research

Ying Huang[1], Wolfgang Glänzel[2], Bart Thijs[3] and Lin Zhang[4,*]

[1]*ying.huang@kuleuven.be*

Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven (Belgium)
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)

[2]*wolfgang.glanzel@kuleuven.be*

Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven (Belgium)

[3]*bart.thijs@kuleuven.be*

Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven (Belgium)

[4] *linzhang1117@whu.edu.cn*

School of Information Management, Wuhan University, Wuhan (China)
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)

Abstract

Due to the advantage of solving complex societal problems and generating meaningful innovations, interdisciplinary research (IDR) has attracted broad attention and extensive support from academia, industry, and government. Most of IDR studies, especially about the quantitative measurement of interdisciplinarity, focus on the perspective of integration. Although impact, as another critical dimension of IDR, has been discussed in recent years, most current studies focus on the impact within the academic research domain, e.g., productivity and citation. Exploring the dynamic evolution of knowledge diffusion, particularly across different disciplinary fields, helps our understanding of IDR and its impact. This paper proposed a novel conceptual framework to investigate the interdisciplinary impact in the broader sense of knowledge diffusion. Beyond the citation count, we care about the citing papers' disciplinary diversity and the cognitive distance between the source paper and citing papers.

Introduction

The complexity and comprehensiveness of significant challenges require solutions developed by fostering effective interdisciplinary collaboration and integrating knowledge from different scientific fields. Under these circumstances, interdisciplinary research (IDR) has attracted wide attention. In order to better explore the IDR, conceptualizing the IDR is the priority task. One of the most broadly-accepted definitions of IDR is set forth in a National Academies' report (National Academy of Sciences et al., 2005). Derived from the above definitions, knowledge integration is the essence of IDR; therefore, in the past few years, most of the studies on IDR, especially about the quantitative measurement of interdisciplinarity, focus on the perspective of integration from ideas (such as concepts and theories), methods (techniques and tools), or data from various fields of knowledge (Porter et al., 2007). Scholars have developed a variety of proxy indicators delivering different insights about the interdisciplinary nature of the object under study, especially from the cognitive perspective (Zhang et al., 2016), the organizational perspective (Zhang et al., 2018), and the policy perspective (Huang et al., 2016).
Impact, as another core dimension of IDR, has also been paid some attention to. In fact, the concept of 'impact' is as complex and multi-faced as 'interdisciplinarity'. The impact can be interpreted from different perspectives. Most of the prior works focus on the impact within the academic research domain, e.g., productivity and citation (Judge et al., 2012; Larivière & Gingras, 2010; Larivière et al., 2015; Steele, 2000; Wang et al., 2015; Yegros-Yegros et al., 2015). In recent work, Zhang et al. (2021) take the broader impact (particularly, PloS usage) into consideration. Although citation impact and 'societal impact' are definitely the important part of the impact of IDR, whether IDR has diffused the knowledge to broader and diverse disciplines are essential dimension deserved to be further investigated.

Method

Forward citation diversity

Forward citation diversity (FCD) is an indicator to measure the diversity among journals citing a given body of research, and "diffusion score" is one of the FCD indicators that is used to trace the influence (Carley & Porter, 2012). Figure 1 presents a concrete example of the relationship between source paper and citing papers. Here, we adopted the full-counting method for the multi-assignment subject system.

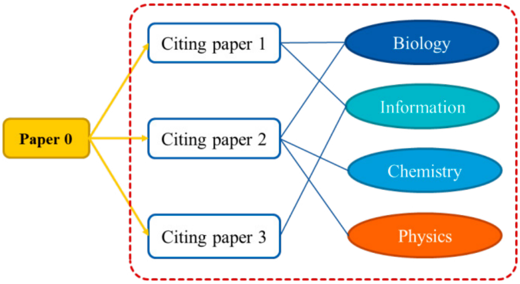


Figure 1. A concrete example of the relation of source paper and citing papers with their corresponding fields

The measurement of FCD is similar to the indicators used for the integration of knowledge generated in different disciplines. It also considers three distinct components: the number of disciplines citing (variety), the distribution of citations among disciplines (balance), and how similar or dissimilar these categories are (disparity) (Stirling, 2007). Its formulate is:

$$FCD{=}1 - \sum_{ij} PF_i * PF_j * SIM_{ij}$$

Where PF_i denotes the proportion of citing publications assigned to the category i in a given paper, SIM_{ij} is Salton's cosine measure of similarity between category i and j. *The* different implementations and comparisons are introduced in our two recent works (Thijs et al., 2021; Huang et al., 2021). FCD scores range from 0 to 1, with higher scores indicative of more significant infiltration into external disciplines. According to the sample similarity/disparity among the four fields provided in Table 1, the FCD value of Paper 0 is 0.478.

Table 1. The sample similarity (disparity) among four fields

	Biology	Information	Chemistry	Physics
Biology	1.0 (0.0)			
Information	0.2 (0.8)	1.0 (0.0)		
Chemistry	0.6 (0.4)	0.1 (0.9)	1.0 (0.0)	
Physics	0.4 (0.6)	0.3 (0.7)	0.8 (0.2)	1.0 (0.0)

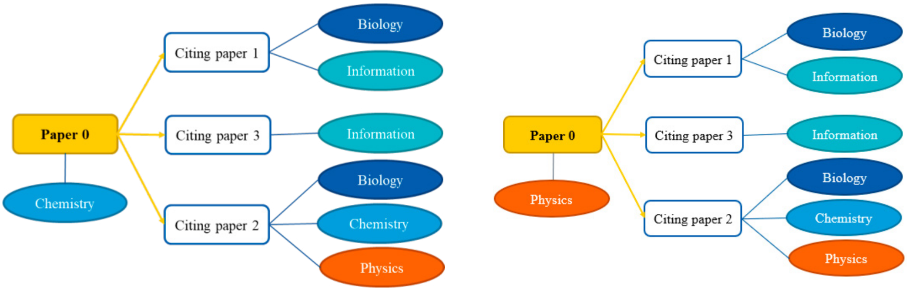
Interdisciplinary impact

The FCD addresses the disciplinary diversity of the citing papers, and it can be treated as an important proxy to trace the knowledge diffusion impact. However, when we consider the source paper's disciplinary category, the situations drive us to figure out supplementary indicators further to cover more different circumstances: a chemistry paper that diffuses the knowledge to other fields is different from a phsics paper that diffuses its knowledge to the same fields. The concrete example of two source papers with different fields and citing papers with their corresponding fields, is shown in Figure 2.

We propose the interdisciplinary impact (IDI) to measure the average disciplinary distance among the source paper and citing papers, and the general formula is as follows:

$$IDI=\frac{1}{n}\sum_{k=0}^{n}\sum_{ij}PC_i * PF_j * DIS_{ij}$$

Where n is the number of citing publications (forward citation), PC_i denotes the proportion of category i in the current publication, PF_j denotes the proportion of category j in the citing publication (forward citation), DIS_{ij} refers to distances between subject field i and j, and $DIS_{ij} = 1 - SIM_{ij}$. With IDI scores range from 0 to 1, higher scores indicate greater knowledge diffusion to distant disciplines.



(a) Source publication is a Chemistry paper

(b) Source publication is a Law paper

Figure 2. Example of two source papers cited in different fields

Forward citation diversity vs. interdisciplinary impact

Either FCD or IDI fails to track interdisciplinary knowledge diffusion of IDR comprehensively; however, combining these two becomes an alternative solution. Figure 3 provides a schematic representation of this twofold perspective, following the conceptualization diagram proposed by Rafols and Meyer (2010). We take individual publications and study knowledge diffusion through their citing papers set, each of the parts in the networks represents a paper (the middle one represents the source paper, and three ones are the citing papers), each shape placed in the oval/round refers to categories of the given paper. Each link indicates the citation between the source paper and citing papers.

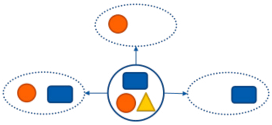
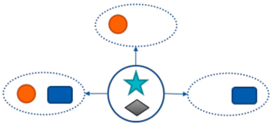
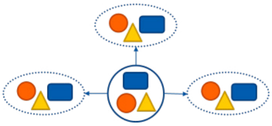
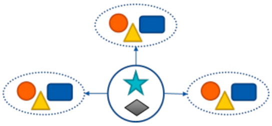
		Interdisciplinary impact (IDI)	
		Low	High
Forward diversity index (FCD)	Low		
	High		

Figure 3. Conceptualization of knowledge diffusion in FCD and IDI

There are four possible combinations:

- High IDI-High FCD: A researcher/study delivers knowledge to distant and diverse fields.
- High IDI-Low FCD: A researcher/study delivers knowledge to distant but limited fields.
- Low IDI-High FCD: A researcher/study delivers knowledge to adjacent but diverse fields.

- Low IDI-Low FCD: A researcher/study delivers knowledge to adjacent and limited fields.

Sample and Results

Based on the indicators proposed in the "Method" section, we compare the results at the level of individual paper level and at the individual researcher level.

Individual paper level

Table 2 presents the results for disciplinary diversity and interdisciplinary impact from the perspective of citing papers for the selected four articles.

Table 2. The knowledge diffusion impact-related indicators of the selected paper

UT	Journal	Num_SC	GE_SC	DIS_SC	FCD	IDI	Type
000275291600016	Chimia	9	0.674	0.830	0.669	0.941	(i)
000315970300193	PLoS One	8	0.422	0.860	0.490	0.794	(ii)
000315307600009	J. R. Soc. Med.	9	0.778	0.813	0.630	0.446	(iii)
000232795300006	Libr. Infor. Sci. Res.	8	0.383	0.704	0.399	0.254	(iv)

Note: Num_SC means the number of subject categories assigned to citing publications; GE_SC indicates the evenness of distribution among citing subject categories; DIS_SC is the average dissimilarity between citing subject categories.

The first paper entitled "How long is the peer review process for journal manuscripts? A case study on Angewandte Chemie International Edition" (000275291600016) published in *Chimia* (Chemistry, Multidisciplinary) is cited in the nine fields: Information Science & Library Science (5), Computer Science, Information Systems (3), Agronomy (1), Computer Science, Interdisciplinary Applications (1), Education & Educational Research (1), Horticulture (1), Medicine, General & Internal (1), Microbiology (1), and Plant Sciences (1). Here, we assume the categories of papers are the same as the classification given to the journal it is published in --it is not perfect but compromised way, and we will discuss it in the "Conclusion and Discussion" section. The first paper delivers information to distant fields, ranging from natural science to life sciences & biomedicine, and social science, and the distribution of citing fields is diverse and relatively balanced.

The second paper entitled "Macro-indicators of citation impacts of six prolific countries: InCites data and the statistical significance of trends" (000315970300193) published in *PLoS One* (Multidisciplinary Science) is cited in the eight fields: Information Science & Library Science (17), Computer Science, Interdisciplinary Applications (12), Computer Science, Information Systems (2), Multidisciplinary Sciences (2), Sociology (1), Surgery (1), Urban Studies (1), and Clinical Neurology (1). The second paper is also cited by the articles from distant fields, but their distributions are so unbalanced that they result in a relatively lower FCD (The citing fields mainly focus on the Information Science & Library Science, and Computer Science, Interdisciplinary Applications).

The third paper entitled "Enhancing the *h* index for the objective assessment of healthcare researcher performance and impact" (000315307600009), published in *Journal of the Royal Society of Medicine* (Medicine, General & Internal), is cited in the nine fields: Medicine, General & Internal (4), Computer Science, Interdisciplinary Applications (1), Information Science & Library Science (1), Education, Scientific Disciplines (1), Emergency Medicine (1), Health Policy & Services (1), Rehabilitation (1), Surgery (1), and Urology & Nephrology (1). Unlike the second one, the third paper is cited by a broad field, but these fields are closer to the field the third paper is assigned to.

The fourth paper entitled "Preference for electronic format of scientific journals - A case study of the Science Library users at the Hebrew University" (000232795300006), published in

Library & Information Science Research (Information Science & Library Science) is cited in the eight fields: Information Science & Library Science (20), Computer Science, Information Systems (5), Biochemistry & Molecular Biology (1), Biology (1), Cell Biology (1), Computer Science, Interdisciplinary Applications (1), Multidisciplinary Sciences (1), and Psychology, Applied (1). On the one hand, the fourth paper is mostly cited by the same field (Information Science & Library Science) or the adjacent field (Computer Science, Information Systems); on the other hand, the distribution of its citing fields are unbalanced.

Individual researcher level

We selected the Derek de Solla Price Memorial Medal winners, who published more than 20 research articles during 2005-2014, as the targeted sample. We first measured the related indicators at the paper level, then aggregated them up to the personal level when relevant by taking the mean score across each researcher's set of papers. The results are shown in Table 3.

Table 3. The knowledge diffusion impact-related indicators of the selected Price award winners

	Aver. Num_SC	Aver. GE_SC	Aver. DIS_SC	FCD	IDI
Bar-Ilan, Judit	9.811	0.645	0.709	0.460	0.421
Bornmann, Lutz	12.638	0.543	0.801	**0.555**	**0.511**
Cronin, Blaise	9.100	0.548	0.715	0.451	0.432
Egghe, Leo	7.193	0.742	0.635	0.405	0.464
Glänzel, Wolfgang	11.203	0.532	0.775	**0.507**	0.477
Leydesdorff, Loet	16.042	0.530	0.797	**0.585**	**0.554**
Moed, Henk F	13.069	0.540	0.788	**0.518**	0.481
Rousseau, Ronald	6.133	0.678	0.649	0.415	0.432
Schubert, Andras	11.821	0.530	0.776	**0.519**	0.467
Thelwall, Mike	11.805	0.590	0.737	0.499	0.478
van Raan, Anthony F J	14.929	0.496	0.813	**0.543**	**0.509**

According to the above table, we can see in the last two columns that Loet Leydesdorff, Lutz Bornmann, and Anthony F J van Raan produce more knowledge diffusion impact on the follow-up studies, so they get a higher score in FCD and IDI. Henk F Moed, Andras Schubert, Wolfgang Glänzel also diffuse their knowledge to a broad range of fields, but these fields are relatively close to the field of the source journal they published, so they perform well in the FCD but relatively lower in IDI.

Conclusion and Discussion

Knowledge diffusion is a dynamic process compared to knowledge integration since citations to a set of published documents can be continuously changing. Therefore, exploring the dynamic evolution of knowledge diffusion, particularly across different disciplinary fields, helps our understanding of IDR and its impact. This paper proposed a novel conceptual framework to investigate the interdisciplinary impact in the broader sense of knowledge diffusion. Compared to the academic impact and social impact mentioned in the previous studies, the combination of FCD and IDI provides a comprehensive scope to understand the broader impact, especially for the IDR that should not only care about the integration of knowledge from two or more disciplines, but also the diffusion of knowledge to solve complex societal problems and generate meaningful innovations. Furthermore, since diffusion can be thought of as integration in reverse, the IDI indicator can also be used to better measure integration--not only considering the disciplinary diversity in the cited references but also the average distance between the cited references and source publication. The IDI is very sensitive to the classification systems of publications. Most of the current classifications are based on journal assignment in which all papers published in a given journal are classified in the same discipline (or set of disciplines) rather than determined by the real content present in the paper. Therefore, a document-based improvement of the subject assignment is necessary. Our other recent study gives solutions for individual-document-based subject assignments by introducing a multiple-generation reference model (Glänzel et al., 2021) .

Acknowledgments

The research underlying this study is done within the framework of the project "Interdisciplinarity & Impact" (2019-2023) funded by the Flemish Government.

References

Thijs, B., Huang, Y. & Glänzel, W. (2021), *Comparing different implementations of similarity for disparity and variety measures in studies on interdisciplinarity*. Proceedings of the 18th Conference of the International Conference on Scientometrics and Informetrics, Leuven (virtual meeting), 12–15 July 2021, in this volume.

Carley, S., & Porter, A. L. (2012). A forward diversity index. *Scientometrics, 90*(2), 407-427.

Glänzel, W., Thijs, B., & Huang, Y. (2021). *Improving the precision of subject assignment for disparity measurement in studies of interdisciplinary research.* Proceedings of the 18th Conference of the International Conference on Scientometrics and Informetrics, Leuven (virtual meeting), 12–15 July 2021, in this volume.

Huang, Y., Glänzel, W., Thijs, B., Porter, A. L., & Zhang, L. (2021). The comparison of various similarity measurement approaches on interdisciplinary indicators. KU Leuven.

Huang, Y., Zhang, Y., Youtie, J., Porter, A. L., & Wang, X. (2016). How does national scientific funding support emerging interdisciplinary research: A comparison study of Big Data research in the US and China. *Plos One, 11*(5), e0154509.

Judge, W. Q., Weber, T., & Muller-Kahle, M. I. (2012). What are the correlates of interdisciplinary research impact? The case of corporate governance research. *Academy of Management Learning & Education, 11*(1), 82-98.

Larivière, V., & Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology, 61*(1), 126-131.

Larivière, V., Haustein, S., & Börner, K. (2015). Long-distance interdisciplinarity leads to higher scientific impact. *Plos One, 10*(3), e0122565.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2005). *Facilitating Interdisciplinary Research*. The National Academies Press.

Porter, A. L., Cohen, A. S., David Roessner, J., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics, 72*(1), 117-147.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics, 82*(2), 263-287.

Steele, T. W. (2000). The impact of interdisciplinary research in the environmental sciences: A forestry case study. *Journal of the American Society for Information Science, 51*(5), 476-484.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707-719.

Wang, J., Thijs, B., & Glänzel, W. (2015). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *Plos One, 10*(5), e0127298.

Yegros-Yegros, A., Rafols, I., & D'Este, P. (2015). Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity. *Plos One, 10*(8), e0135095.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology, 67*(5), 1257-1265.

Zhang, L., Sun, B., Chinchilla-Rodríguez, Z., Chen, L., & Huang, Y. (2018). Interdisciplinarity and collaboration: on the relationship between disciplinary diversity in departmental affiliations and reference lists. *Scientometrics, 117*(1), 271-291.

Zhang, L., Sun, B., Lidan, J., & Huang, Y. (2021). On the relationship between interdisciplinarity and impact: Distinct effects on academic and broader impact. *Research Evaluation.* https://doi.org/10.1093/reseval/rvab007

Where to start reading? Introducing the Reference-Citation Plot

Marcus John[1], Frank Fritsche[2] and Christian Gülden[3]

[1] *Marcus.John@int.fraunhofer.de*
Fraunhofer Institute for Technological Trend Analysis INT, Appelsgarten 2, D 53879 Euskirchen (Germany)

[2] *Frank.Fritsche@int.fraunhofer.de*
Fraunhofer Institute for Technological Trend Analysis INT, Appelsgarten 2, D 53879 Euskirchen (Germany)

[3] *Christian.Guelden@int.fraunhofer.de*
Fraunhofer Institute for Technological Trend Analysis INT, Appelsgarten 2, D 53879 Euskirchen (Germany)

Abstract

With this contribution we address the problem of identifying so called key publications, i.e. publications, one should read, in order to generate an overview of a topic. To this end we introduce the reference citation plot as a useful approach for identifying such publications by plotting a paper's number of references and number of citations, respectively. The first one was chosen in order to estimate the broadness of a publication's knowledge base. This concept is perpetuated by introducing the Reference and the Citation Topic Width, which measure the disciplinary broadness of the references and of the citing papers. We demonstrate the usage of these metrics within the field of technology foresight.

Introduction and problem

Where should one start reading, if one delves into a scientific topic? What are the publications one should definitely read? This is one of many problems, scientists and researchers have to deal with in their every-day work. Within technology foresight, this problem is even aggravated, since in many cases futurist have to deal with very different topics stemming from various scientific fields, although they are no experts in this particular field (Schulze et al. 2020). An obvious approach to address this problem would be, to start with a suitable review paper. However, it has been shown by Harzing (2013) and Colebunders & Rousseau (2013), that a significant number of publications in the Web of Science might have been misclassified regarding their document type. This observation is consistent with our experience in carrying out literature research on a regular basis. Thus, we wanted to establish a plausible and yet feasible paper-based approach, which should be easy to implement, in order to identify *key publications*. Those are publications, which are particularly suitable to gain an overview of a topic and to identify current research highlights. Thus, the notion of a key publication is broader than that of a review paper as it tries to take into account the different aspects of a typical foresight process.

Our approach

Description of the system used

We used an in-house database based on the Web of Science data, which is part of a support system for technology foresight, which has been developed at the Fraunhofer INT (John, 2018). The system consists of three components. First, a graph database implementing an RDF graph model forms the core of the system. This approach is particularly well-suited for storing highly interconnected bibliographic data by utilizing a generic data model. One advantage of this approach is, that it allows to reduce complex analyses to simple queries using SPARQL, the corresponding query language for RDF graph databases. A high-performance search engine is used to perform searches on the approximately 65 Million publications (as of January 2021) stored in the database. The technology analysts at the Fraunhofer INT interact with the systems via an intuitive graphical user interface, which consists of two different parts, namely a search interface for a typical literature research, and an analytic view, including different dashboards. The latter provides a comprehensive data-driven view on a topic and the corresponding publications.

Description of the metrics considered

In order to identify those publications, which provide a suitable overview over a specific topic, we make the following assumption: an author must have acquired a rather broad knowledge base to be able to present such an overview. In contrast, papers which deal with certain specific and detailed aspects of a topic, do not necessarily refer to the aggregated body of knowledge on this topic. As a consequence, we wanted to construct a proxy for measuring the broadness of a paper's knowledge base. Therefore, we considered the number of references N_{ref} of a paper as first proxy to measure the broadness of its knowledge base. Since citing and being cited only depends on the point of view, it seemed logical to use the number of citations T_c as a second metric. Thus, we implemented a simple scatter plot in our system, which depicts N_{ref} on the x-axis and T_c on the y-axis. We called this plot *reference-citation plot,* and it is shown in Figure 2.

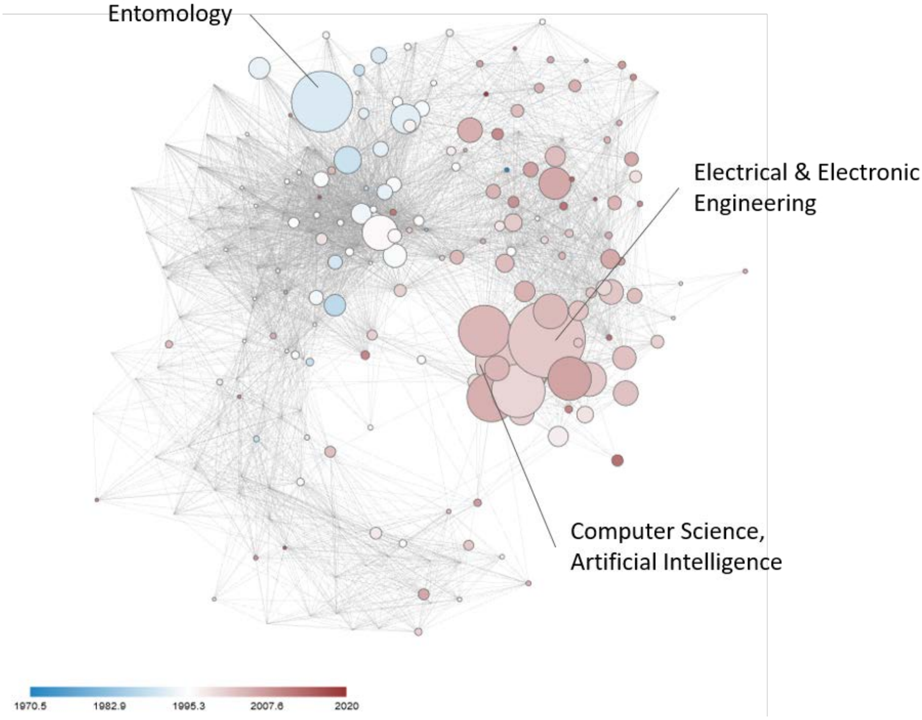


Figure 1: Overlay Map for generating an overview of the topic "bee colony". Each bubble represents a subject category and its color is determined by the mean age of publications in the respective category.

In a second step we measured the number of disciplines, indicated by the subject categories of the Web of Science, the references originated from. We call this the *Reference Topic Width RTW*. It is based on the observation, that some papers access knowledge from only a few different disciplines, while others have a broad knowledge base meaning they are referring to publications from many disciplines. Finally, we once again refer to the mirror-inverted relationship between citing and being cited and introduce the *Citing Topic Width CTW* which represents the number of disciplines the citing papers of a publication are assigned to.

A didactic example

As an example, we chose the topic “bee colony”. The reason for this will become clear at the end of this contribution. Using the simple search query “bee colon*”, which corresponds to a typical topic search in the Web of Science web-interface, resulting in 8.253 publications as of January 2021.

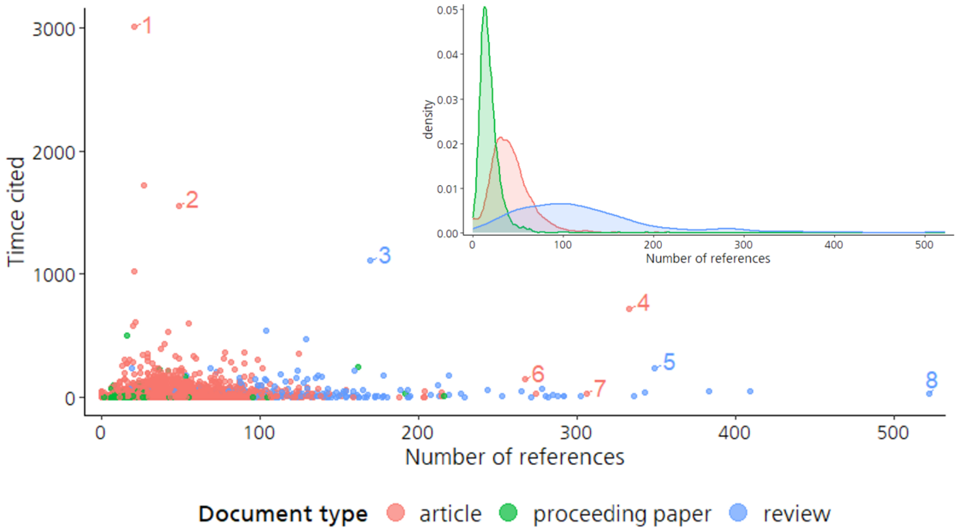


Figure 2: Reference Citation Plot for the topic “bee colony”. Details for the eight marked papers are given in Table 1. The inlay displays a density plot for the distribution function of the number of references for three different document types.

Generating an overview

First, we present an overview over the topic by using the overlay maps developed by Rafols et al. (2010). This kind of representation of the disciplinary structure of a topic was extended by adding further options for coloring the bubbles, which represent the subject categories. The possibility to scale the color by the mean age of the publications (see Figure 1) proved to be particularly useful. This coloring scheme reveals, that the topic “bee colony” is obviously partitioned between two different broader research areas. The first and older one, as indicated by the blue colors, can be found in the upper left and comprises subject categories stemming from the life sciences, like Entomology, Zoology, Ecology etc. The second, broader area covers disciplines from Computer Science and Engineering. These disciplines contain more recent publications as indicated by the red colors of the bubbles. This example demonstrates, that coloring the bubbles offers the opportunity to gain further and deeper insight into the disciplinary structure of a topic which is particularly valuable in the context of technology foresight. Since the user is enabled to interact with the map in the browser-based system, it still further enhances the opportunities of such an analysis.

The Reference Citation Plot

The reference citation plot for the topic “bee colony” is shown in Figure 2. The underlying idea of this representation is to generate a quick overview. Possible survey articles can be found in the lower right part of the plot, regardless of whether they are tagged correctly as a review paper or not. As can be seen, there is a considerable number of articles with more than 100 references which are not tagged as review papers. Indeed, the titles of the paper 4, 6 and 7 (see Table 1) confirm, that these papers have the character of a review, although they are labelled as articles. In order to check, whether our assumption regarding the amount of references of review papers is correct, we analyzed the frequency distribution of this number using a density plot (see Figure 2, inlay). The observed shift of the maximum of the distribution confirms our assumption. As expected proceeding papers tend to have fewer references (maximum at approximately 14 references), than articles (maximum at ~35) and review papers (maximum at ~100).

Table 1. Title and key indicators of the eight publications marked in Figure 2 and 4. The color of the title indicates the document type, i.e. red for articles and blue for review papers.

Title	*Number*	*TC*	*Nref*	*RTW*	*CTW*
A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm	1	3008	21	7	115
A comparative study of Artificial Bee Colony algorithm	2	1556	49	16	96
Bee declines driven by combined stress from parasites, pesticides, and lack of flowers	3	1107	170	29	78
A comprehensive survey: artificial bee colony (ABC) algorithm and applications	4	718	333	48	83
A Comprehensive Survey on Particle Swarm Optimization Algorithm and Its Applications	5	239	349	74	95
A survey of swarm intelligence for dynamic optimization: Algorithms and applications	6	144	267	33	34
Multi-population techniques in nature inspired optimization algorithms: A comprehensive survey	7	27	306	58	22
Catalog of the Indo-Malayan/Australasian stingless bees (Hymenoptera: Apidae: Meliponini)	8	25	522	23	12

A closer look at the titles also reveals the reason for the thematic split observed in the overlay map in Figure 1. While, as expected, the biological part deals with bees (e.g. paper 3 and 8 in Table 1), the computational and engineering part is about a global optimization algorithm named artificial bee colony (ABC) algorithm.

Additionally, the reference citation plot allows the quick identification of highly cited papers, which are located in the upper left part of the plot. Such papers are possible breakthrough papers, which might indicate important developments within a topic. Paper 1 in Figure 2 is an example for this hypothesis, as it described the aforementioned optimization algorithm for the first time.

Utilizing the disciplinary broadness of references and citations

In Figure 3 we present a scatter plot of the RTW and the CTW. Again, we marked the papers listed in Table 1 for a rough orientation. The first intriguing observation is, that in some cases review articles tend to have a broader knowledge base in terms of disciplines than regular articles (e.g. paper 4, 5 and 8 in Table 1). Other, rather thematically specific publications, like paper 3 or 8, build upon a knowledge base, which comprises only a few disciplines, resulting in a low value of the RTW. From the point of view of technology foresight this distinction is

quite useful, since it allows researchers to select the necessary literature for a deep dive into a topic.
The examples given in Table 1 regarding the CTW demonstrate, that this metric is useful too, since it measures the disciplinary broadness of a publication's impact. This becomes particularly clear with respect to paper 1. Although it has a rather narrow knowledge base as measured by RTW, it has influenced applications in many different areas as can be concluded from the high value of the CTW. We were able to confirm this, by taking a closer look at the overlay map, where we identified a number of applications of the ABC algorithm stemming from different disciplines and fields of application. We want to point out, that this is an important insight for technology foresight.

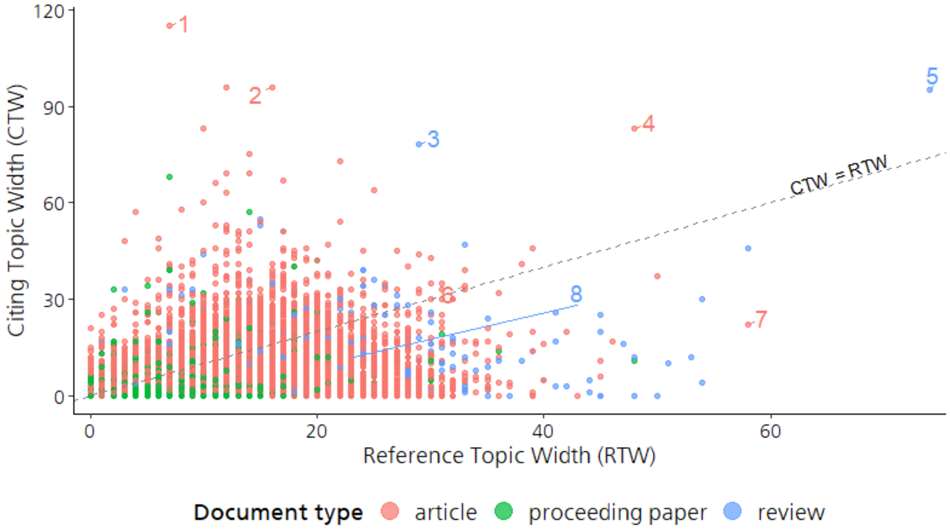


Figure 3: Scatter plot displaying RTW and CTW. Additionally, the line where RTW equals CTW is shown. Publications below this line have a larger RTW than CTW.

Furthermore, in some cases we observed, that the disciplinary perception of a paper differs significantly from the disciplinary structure of its knowledge base, which is also apparent from Figure 3. In order to analyse this aspect, we first took a closer look at the difference between CTW and RTW. If this difference is greater than 0, the paper's disciplinary perception is broader than its knowledge base. A density plot of the distribution of this difference is presented in Figure 3. Although in most cases, the difference is smaller than 0, indicating a disciplinary narrowing, there are some publications for which we observed a disciplinary broadening. Such a shift is of possible interest for technology foresight and needs to be investigated further.

Conclusion & Outlook

As demonstrated, the idea to measure the broadness of a paper's knowledge base is intriguing and helpful for identifying key publications. Although we can only present a simple anecdotic example here, the approach has proven quite useful in everyday literature research at the Fraunhofer INT. Of course, a more systematic investigation is necessary in order to overcome its anecdotical character. It is noteworthy, that this approach is transferable to other databases.

The difference of the CTW and RTW values indicate, that the disciplinary structure of a paper's knowledge base and its citing papers have changed. Such a shift might be an important indicator, which might be particularly useful in the context of technology foresight.

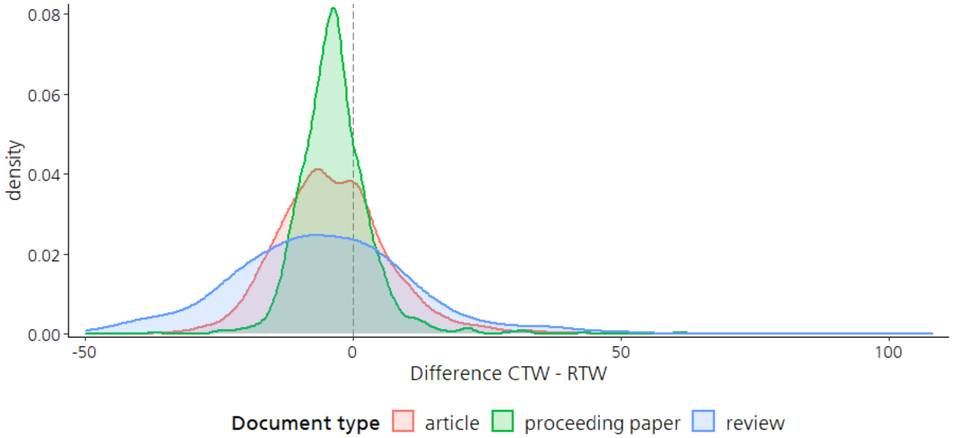


Figure 4: Density plot of the difference between CTW and RTW.

There are several directions of future research. First of all, not only the number of different disciplines should be considered, since it only covers the variety of the knowledge base (Rafols & Meyer, 2010). It could prove useful, to assess its diversity too. Additionally, different categorization systems such as those implemented in the Dimensions system (Herzog, Hook, & Konkiel, 2020) should be explored.

References

Colebunders, R., & Rousseau, R. (2013). On the Definition of a Review, and does it matter? In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. F. Moed (Eds.), *Proceedings of ISSI 2013 Vienna* (pp. 2072–2074). Vienna: AIT - Austrian Inst. of Technology.

Harzing, A.-W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences? *Scientometrics, 94*(1), 23–34.

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies, 1*(1), 387–395.

John, M. (2018). Data driven foresight - Technologiefrühaufklärung im Zeitalter von Big and Linked Data. Ein Werkstattbericht. In J. Gausemeier, W. Bauer, & R. Dumitrescu (Eds.): *Vol. 385. HNI-Verlagsschriftenreihe, Vorausschau und Technologieplanung* (pp. 409–421). Paderborn.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics, 82*(2), 263–287, from doi:10.1007/s11192-009-0041-y.

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology, 61*(9), 1871–1887.

Schulze, J., Grüne, M., John, M., Neupert, U., & Thorleuchter, D. (2020). Futures Research as an Opportunity for Innovation in Verification Technologies. In I. Niemeyer, M. Dreicer, & G. Stein (Eds.), *Nuclear non-proliferation and arms control verification. Innovative systems concepts* (pp. 187–204). Cham: Springer.

Reference Distributions of Reviews in the Web of Science – The Effects of Company Policy from 1992-2010

Miloš Jovanović[1], Philipp Baaden[1] and Frank Fritsche[1]

[1] *milos.jovanovic@int.fraunhofer.de*
philipp.baaden@int.fraunhofer.de
frank.fritsche@int.fraunhofer.de
[1]Fraunhofer Institute for Technological Trend Analysis INT, Appelsgarten 2, 53879 Euskirchen (Germany)

Abstract

Scientific literature reviews are an important entry point for researchers to get a first overview of a scientific topic and to find more publications that are worth reading. Also, evidence exists that reviews are more often cited than original research articles. This makes it important that reviews are correctly classified in literature databases. In our research, we used the Web of Science (WoS) by Clarivate Analytics and analysed a total of 1.8 million reviews. Clarivate had a review classification policy in force from 1992-2010. In the present paper, we wish to highlight the effects this policy has had on the distribution of reviews in the WoS in general and also in the Subject Categories. We find distinct distributions in groups of Subject Categories, discuss their meaning and potential pitfalls for researchers wanting to use reviews in the WoS.

Introduction

Dedicated scientific literature reviews (henceforth referred to in this paper as "reviews") constitute a central element of scholarly communication. Reviews strive to collect scientific knowledge in the form of research papers and present them in a systematic and orderly way. As such, a review also represents a selection of research papers that in the opinion of the authors represents the most important and current research on different topics. Normally, reviews do not describe new research themselves but rather the research contained in other research articles. In this way, reviews are an important entry point for researchers to get a first overview of a scientific topic and to find more publications that are worth reading. The University of Texas Libraries describes them as "virtual gold mines if you want to find out what the key articles are for a given topic" and "review articles are good places to get a basic idea about a topic" (see references for the website address. Reviews are also written for scientists in their own respective communities so that they can look back at recent developments of their field. At our institute where we conduct many studies in the field of technological foresight, reviews are an important aspect for our technology analysts to get a first understanding of current research on different topics (Torraco 2016).

According to Colebunders und Rousseau (2013) and Colebunders et al. (2014) the number of reviews in the fields of Tropical Medicine, Infectious Disease and Oncology increased and they were more likely to be cited than were normal research articles. Since, as these authors note, "impact factors and citations are increasingly used to evaluate scientists and […] reviews receive more citations than do normal research articles" an interesting question to answer is how high the number of "mislabeled reviews" in the WoS is, since reviews might be cited which actually are no reviews and actual reviews cannot be found because they are not classified as reviews.

In her excellent paper on the same topic, Harzing (2013) describes the situation she found in 2011. She analysed the document categories with regard to the Social Sciences (also mentioning the Sciences) and used a data set based on a journal sample. The present research-in-progress paper wishes to build upon this work and expand the data set used and the scope.

The WoS describes a very straightforward definition of what a review is. On one of the official websites by Clarivate Analytics (the company behind the WoS) the following definition is listed (Garfield 1994):

> *"In the JCR system any article containing more than 100 references is coded as a review. Articles in "review" sections of research or clinical journals are also coded as reviews, as are articles whose titles contain the word "review" or "overview.""*

It is important to note a couple of things about this definition, though. The text on the Clarivate site (https://clarivate.com/webofsciencegroup/essays/impact-factor/) was originally published by Eugene Garfield in a comment in the "Current Contents", a source of information on different journals. The present authors were unable to get a copy of the original paper, but there are at least two digital versions of them, one of those the text on the Clarivate homepage. Interestingly enough the text on the homepage is edited, at least because of the insertion of the name "Clarivate Analytics" a company that was only founded in 2016. Whether or not more changes to the original text have been made could not be ascertained. It is also important that Garfield notes that reviews are often more cited than "normal" research articles but the policy he describes might lead to artificial inflation of the number of reviews inside of a journal. The other important thing is that the text is published under a section called "essays", and as such does not constitute the official policy of the WoS. However, the described procedure was used for a period of time, as also Harzing (2013) describes in her paper. It is this policy of the WoS and its effect on document classification that we study in detail in the present paper.

Reaching out to Clarivate Analytics, we got the information via personal email correspondence that indeed such a policy was in force from 1992-2010 and that the classification of a review was based upon a) the publisher if he lists an article under a section that implies that it is a review, b) the words "Review of the literature" in the title of the article, the abstract or the introduction, c) the words "Review, Systematic Review or Mini-review" appear in the title of the article and must appear elsewhere in the article, d) the article in question has more than 100 references and finally e) reviews which were presented at conferences were processed as Proceedings Papers.

Among others, the central questions we wanted to answer was what effects this policy had on the reviews in the WoS in general and whether it affected the different scientific disciplines (based on the so-called Subject Categories (SCs) of the WoS) in different ways. We used these categories since they were readily available in the data.

Methods

Our institute uses both the online version of the WoS and an offline database of the same data. The latter allows us to employ more detailed analyses than with the online tools alone. We are aware that other institutions might have such an offline access to the data and thus other options to find reviews or review-like papers. It is our opinion, though, that for most researchers, the online version is the one typically used. Our data subscription includes the Science Citation Index Expanded and the Social Sciences Citation Index for the years 1970 till today as well as the Conference Proceedings Citation Index – Science and the Conference Proceedings Citation Index – Social Science & Humanities for the years 1990 till today. In total, our offline database includes about 62 million publications (in June 2020).

For the present study, we used a data set from July/August 2020 containing all publications that have the document type "Review". In total, this data set contains ca. 1,8 million reviews. We also split this data set by SCs since those are an approximation of scientific disciplines. We did this in order to see whether there were any differences between reviews in, e.g. the life sciences or engineering. To analyse the data sets, we used RStudio and different scripts in R. The analytical approach is straightforward. The data set contains information about the WoS-ID of the publication, its publication year, the total number of references, and finally the SCs to which the paper is assigned. It should be noted that a paper can be assigned to multiple SCs.

Results

As a first approach to the data set, we took a look at all publications with the document type "Review" (1.8 million publications) and sorted them by the number of references each had. In a next step, we deleted the reviews with 0 references (nref = 0) and those with more than 500 (nref > 500). The reason for removing the reviews with zero references (about 30.000 publications) was that from our point of view those did not represent typical scientific reviews, i.e. an overview of the literature of a specific scientific field or topic. Why such publications exist at all seems to be due to at least two reasons. The obvious one is incorrect indexing. For example, a review named "Matrix Metalloproteinases – a review" (WoS-Accession Number: A1993KP25900004) was indexed correctly as a review. However, no cited references were indexed. But looking into the actual PDF of the paper (or, if available, the original article itself) one can see that the publication includes a long list of references (H. Birkedal-Hansen et al. 1993). There are other, current, examples like "An overview of wild edible fruits of Western Ghats, India" by Sreekumar et al. (2020). This paper is also indexed as being "Early Access" so maybe the error in the number of references will be corrected in the future. It seems to us that this is a simple error in indexing and we take note of it. This means that for our analysis, we lost some relevant reviews when we deleted reviews with a nref of 0. This is an error that can be found on the online version of the WoS as well as in our offline data. We did not have the resources to check all excluded reviews manually. We assume this kind of error is responsible for most of the reviews with zero references. However, there can also be found publications that indeed have zero references like Thomas et al. (2011) "The Dawn Spacecraft" which is a description of the titular spacecraft and its mission. In total, we think that with only 2% of all reviews, deleting these papers from our data set is an acceptable trade-off to improve the validity of our data set. As for the reviews with more than 500 references, we found 7673. We excluded those because they represent the very long endtail of an already very long tail in the data set. Including these does not improve the validity of the data set. In total we deleted about 40.000 publications or 2% of all reviews in the Web of Science)

With the remaining 98% of all reviews in the WoS we plotted Figure 1. It can be seen quite plainly that the number of reviews rises with the number of the references until it reaches a peak at about 50 references (with 17.000 reviews having this number of references). After that the number of reviews gradually declines. Thus far, this seems to correspond with the nature of reviews, meaning that a certain amount of referenced literature is needed to actually make a review and after reaching a certain amount of references the number of reviews simply drops off, probably for a variety of reasons (e.g. the amount of work one needs to put into a review grows with each reference). The number of these references probably also depends on the discipline for which the review is written (more on that later).

The more interesting to note in Figure 1 is an unexpected second peak at about 100 references (with ca. 13 000 reviews). This peak is, of course, due to the policy, the WoS used until 2010 to classify any publication as a review which had 100 or more references. But this peak also shows that the number of reviews has thus been artificially inflated, due to company policy. As Harzing (2013) noted that this led to a huge number of normal articles to be classified as reviews. Since the policy was abandoned, Clarivate Analytics was probably also of the opinion that their policy did not match the reality of the scientific publication practices. But since the peak is still there in the data set, what does this mean for researchers, bibliometricians, policy makers and other users of the WoS?

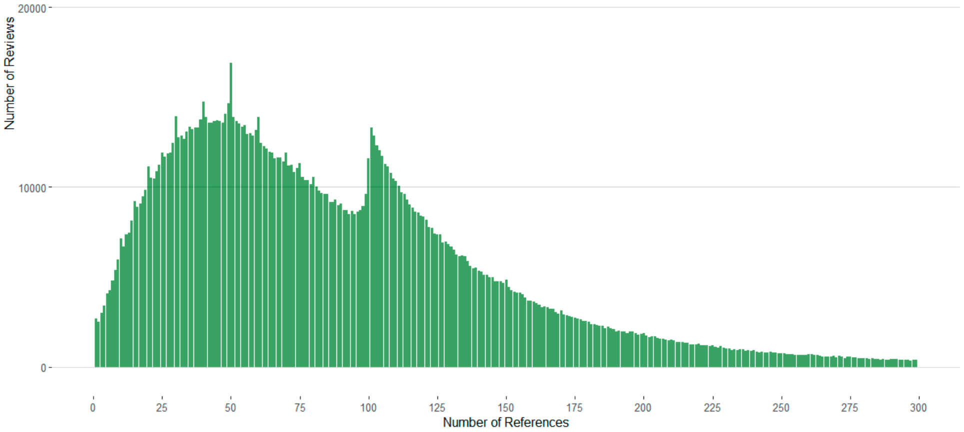


Figure 1: The distribution of reviews in the Web of Science (1970-2019).

Similar to Figure 1 we created graphs for each of the Web of Science’s so-called SCs. SCs are classifications which are used to label every journal in the WoS using at least one SC (some have more). It is important to note that the SCs are journal- and not article-based. This means that it is not always clear which publication from a journal with two or more SCs is part of one, the other, or both SCs. There is a whole body of literature on the question whether the SCs are a good classification of scientific disciplines for journals. See for example Leydesdorff und Rafols 2009, who argue that even though there exist “inaccuracies in the attribution of journals to the ISI SCs” those “presumably […] average out” for the main structure of the disciplines.

Looking at a total of 236 SCs we found that it was possible to categorize the SCs into three broad groups and that two of those did not correspond to the distribution found in Figure 1. Namely, one of the broad groups included most Social Sciences and the Humanities and looked like the examples in Figure 2.

This group has two peaks, one at about nref > 25 and one at about 100. We called this group the “Camel” group because it included two humps.

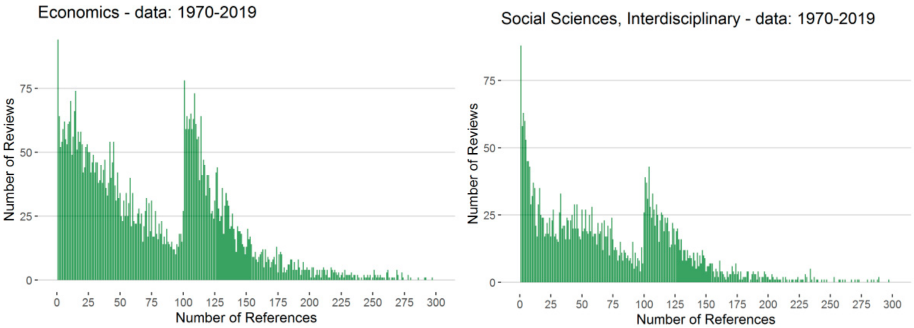


Figure 2: Examples for the „Camel“-group.

The next group we have encountered in our data seems to have a typical distribution that looks like this (see Figure 3):

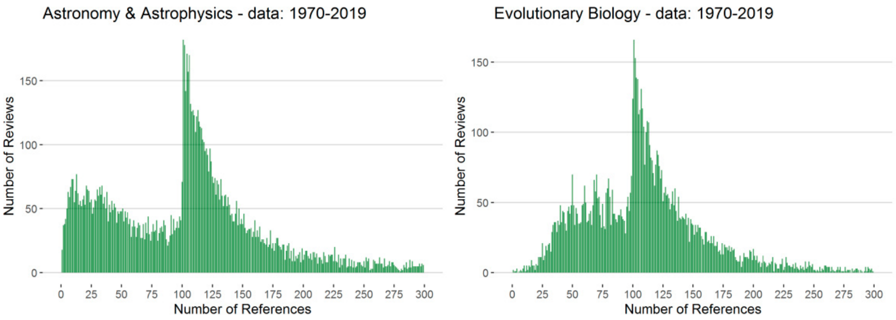


Figure 3: Examples for the „Sharkfin“-group

This kind of distribution has different kinds of peaks left of nref = 100 but at about this point there is a very strong solitary peak, which is why we called this group the „sharkfins“. As to the SCs in which this distribution surfaces, there seems to be no common characteristic but it can rather be found in the natural sciences, engineering, arts and humanities.

The final group we found in our data set had rather typical distributions when looking at bibliographic data (see Figure 4):

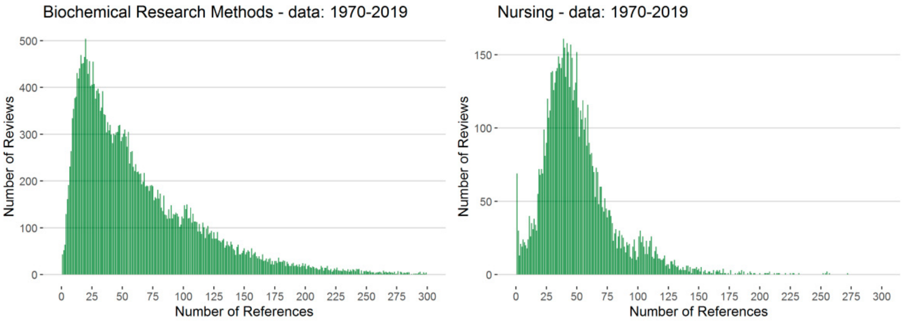


Figure 4: Examples for the left-skewed group of SC-distributions

These distributions seem to follow a Poisson distribution and appear to be typical for medical and life science SCs in general.

To sum up, our results suggest that there are three groups of distributions of reviews:

1. The “Camel”-group that seems to be typical for the Social Sciences and Arts & Humanities.
2. The left-skewed group that is typical for life science SCs.
3. Finally, the “Sharkfins” that can be found in a great variety of SCs.

Discussion and conclusion

In the present research-in-progress paper, we were able to show that Clarivate’s policy from 1992-2010 had a profound effect on the distributions of reviews in the WoS, similar to Harzing (2013) but on a broader scale. Also, according to the distributions, there are groups of SCs that are more strongly affected (the “Camels” and “Sharkfins”) than others (the life science SCs). With this information researchers using the WoS already have an idea which SCs have a higher number of mislabeled reviews than others. In our further research on this topic we will split the data set into three time periods (pre-policy, policy and post-policy) to see how the distribution of reviews in the WoS behaved before and after the policy and whether those had a lower amount of mislabeled reviews. Based on this, we want to calculate an indicator based on the average number of references for reviews of specific SCs in order to recommend possible future policies that keep the amount of mislabeled reviews at a minimum while not being to complicated to implement.

Acknowledgments

We would like to thank the two anonymous reviewers for their helpful and constructive comments which we integrated into the final version of our paper.

References

Birkedal-Hansen H., Moore W.G.I., Bodden M.K., Windsor L.J., Birkedal-Hansen B., DeCarlo A., & and Engler J.A. (1993). Matrix Metalloproteinases: A Review. *Critical Reviews in Oral Biology and Medicine*. (4), 197–250.

Colebunders, R., Kenyon, C., & Rousseau, R. (2014). Increase in numbers and proportions of review articles in Tropical Medicine, Infectious Diseases, and oncology. *Journal of the Association for Information Science and Technology, 65*(1), 201–205.

Colebunders, Robert; Rousseau, Ronald (2013): On the definition of a review, and does it matter? In: *Proceedings of ISSI 2013 - 14th International Society of Scientometrics and Informetrics Conference*, Vol. 2, 2072–2074

Garfield, E. (1994). The Impact Factor. *Current Contents*. (25), from https://clarivate.com/webofsciencegroup/essays/impact-factor/.

Harzing, A.-W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences? *Scientometrics, 94*(1), 23–34

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348–362.

Sreekumar, V. B., Sreejith, K. A., Hareesh, V. S., & Sanil, M. S. (2020). An overview of wild edible fruits of Western Ghats, India. *Genetic Resources and Crop Evolution, 67*(7), 1659–1693.

Thomas, V. C., Makowski, J. M., Brown, G. M., McCarthy, J. F., Bruno, D., Cardoso, J. C., et al. (2011). The Dawn Spacecraft. *Space Science Reviews, 163*(1-4), 175–249.

Torraco, R. J. (2016). Writing Integrative Literature Reviews. *Human Resource Development Review, 15*(4), 404–428.

University of Texas Libraries: Finding Journal Articles 101. Review Article. *University of Texas Libraries.* Online available at https://guides.lib.utexas.edu/c.php?g=556926&p=3829927

Detecting Global Publication Trends in Research Integrity and Research Ethics (RIRE) through Bibliometrics Analysis

K. A. Khor[1], L. G. Yu[1] and A. M. Soehartono[1]

[1] *mkakhor@ntu.edu.sg; mlgyu@ntu.edu.sg; asoehartono@ntu.edu.sg*
Talent Recruitment and Career Support (TRACS) Office and Bibliometrics Analysis, Nanyang Technological University, 76 Nanyang Drive, Singapore 637331 (Singapore)

Abstract

Global publication trends in research integrity and research ethics (RIRE) have increased exponentially. Over 25,700 publications on RIRE from Web of Science across the period 1990-2018 form the basis of this study. Countries, institutions, and their collaborations on RIRE are analyzed. Results show that developed countries in North America and Western Europe contribute to most of the global publications on RIRE. However, despite China's significant growth on RIRE study in recent years, the increase is not comparable to its rapid growth in overall publications for science and technology. International collaboration on RIRE increased from ~2% of the total publications in 1990 to 23% in 2018. Developed countries dominate international collaborations on RIRE, with USA– Canada, USA-UK and UK-Australia as top collaborating pairs. The proportion of multi-authored, multi-institutional and multi-country papers has grown over the years, indicating the trend towards increased collaborations in RIRE studies.

Introduction

Globally, research has been on an upward trajectory. The number of scientific publications has increased exponentially (Larsen and von Ins 2010; Powell et al. 2017; Zhang et al. 2015), fueled by societal and commercial interests. These are evidenced by increases of R&D funding, numbers of researchers, commercialization potential, and participation of emerging countries (Bhattacharya et al. 2015; Javed and Liu 2018; Moiwo and Tao 2013; Shashnov and Kotsemir 2018). Alongside the expanding research landscape, the role of ethics and integrity continues to serve as a basis for good research practices. The necessity and shared responsibility of all stakeholders to uphold standards and principles is expected to continue to grow as more countries commit to funding research and more researchers contribute to the knowledge base (Titus et al. 2008).

Whereas research ethics defines the standard framework of acceptable conduct in research (Resnik 2020), research integrity represents the adherence of conduct to ethical principles which renders trust and confidence of the given methods and results. Both are complementary in supporting responsible scientific conduct (Bird 2006; Braun et al. 2020). The absence of either facet can have potentially dire consequences which could erode public trust in research. Clinical research without ethical protocols can marginalize vulnerable study participants (Schüklenk 2000). Flippant or perfunctory conduct, skewed or exaggerated data, or outright fraudulent results has the potential to misdirect future paths of research (L. M. Bouter et al. 2016; Coxon et al. 2019). Poor research may also mislead or misinform policymakers, thereby undermining government support for research in terms of its reliability and future funding (Michalek et al. 2010). In the competitive research landscape, researchers and institutions also face increasing pressure and demands (Haven et al. 2019), potentially leading to corner-cutting or compromising of standards.

The ongoing discourse led to flourishing of research ethics and research integrity (RIRE)-related fields (Aubert Bonn and Pinxten 2019). In the last 10 to 15 years, the number of retractions has increased (Steen et al. 2013), signalling a greater awareness and identification of appropriate publication conduct (Fanelli 2013). Moreover, institutions are adopting policies and regulations to promote good practices (Geller et al. 2010; Lins and Carvalho 2014). In this work, we explore and identify the global trends of the broad scope of responsible research, in terms of RIRE, through bibliometric analysis, between 1990-2018. We aim to characterize the shifts, trends, and participants in order to provide a broad-view understanding about the research activities and developments over the years.

Data Set and Methods

Web of Science (WoS) was used as database source for the study. The period of study was restricted between 1990-2018. In view of the broad range of dissemination methods in the field, where key papers can take the form of editorials or notes (Aubert Bonn and Pinxten 2019), all publication types were included. Publications containing keywords related to RIRE, such as "scientific ethics" or "scientific integrity" or "publication ethics" or "publication integrity" or "scientific misconduct" or "scientific dishonesty" or "publication misconduct" or "publication dishonesty" or "academic cheating" or "scientific cheating" or "research cheating" or "publication cheating" or “bioethics” or “plagiarism” or "research misconduct" or "research dishonesty" or "academic misconduct" or "academic dishonesty" or "Research Integrity" or "Research Ethics" or "Academic Integrity" or "Academic Ethics" were used for the search. The terms were refined and selected to ensure inclusion of relevant RIRE topics. Over 25,700 publications were identified and used for subsequent bibliometric analysis and citation mapping. For bibliometric analysis, frequency counts were determined based on the field element. To calculate ratios of publications of a country in terms of global shares, publications associated with the target country in each category (RIRE or All Subject Areas) were obtained. The % Share of Global Research Integry/Ethics papers is computed as the ratio of papers associated with an authoring country over the total number of papers in the subject globally. The % Share of Global Research in All Subjects was taken as the publication count per country over the counts on All Subject Areas globally. Data for the latter was obtained from the InCites database. The Share Difference is computed as the difference of % Share of Global Research on RIRE papers and % Share of All Papers on All Subject Areas.

VOSViewer (VOSviewer 2020), developed by CWTS in Leiden University (The Netherlands), was employed to conduct scientific mapping and extract keywords in the publications to elucidate the topics in RIRE. In order to understand the country-specific topics on RIRE, publications on research integrity/ethics were also analyzed by country.

Results and Discussion

Global yearly publications on RIRE between 1990-2018 show a steady increase (Figure 1(a)). Publication volume rose from 101 publications in 1990 to 2315 in 2018, with a compound annual growth rate (CAGR) of 12.8%. The United States dominates in publishing topics related to RIRE with 8594 cumulative publications, followed by England, Canada, Australia, Spain, Brazil, Germany, Italy, France, and China. The majority of top 10 countries researching on RIRE are generally higher-income western nations, with a track-record of reputable institutions and research. China is the only Asian country amongst the top 10 (10th), while India is in the 13th position (Figure 1(b)).

Figure 2 gives the country-level differences between global publication shares on RIRE to shares in All Research Areas for leading countries between 1990-2018.

USA, Canada, UK, and Australia are found to have positive variance, while China, Germany, France, Italy and India have negative variance, indicating that the latter group of countries has relatively less research activity on RIRE compared to overall research activity. Yet, the reasons may differ amongst countries. The lag in RIRE-related research within the emerging countries is thought to be attributable to a host of factors, such as publication priorities in technical disciplines, or ongoing development and enforcement of research ethics and integrity policies within the institutions (Ana et al. 2013; Fanelli et al. 2015; Okonta and Rossouw 2014). For example, China exhibited exceptional publication growth, accounting for 8.3% of all global research publications between 1990-2018 in all subject areas (the second country by scholarly output after the United States). By contrast, publications on RIRE only accounted for 2.0% of the global output on RIRE.

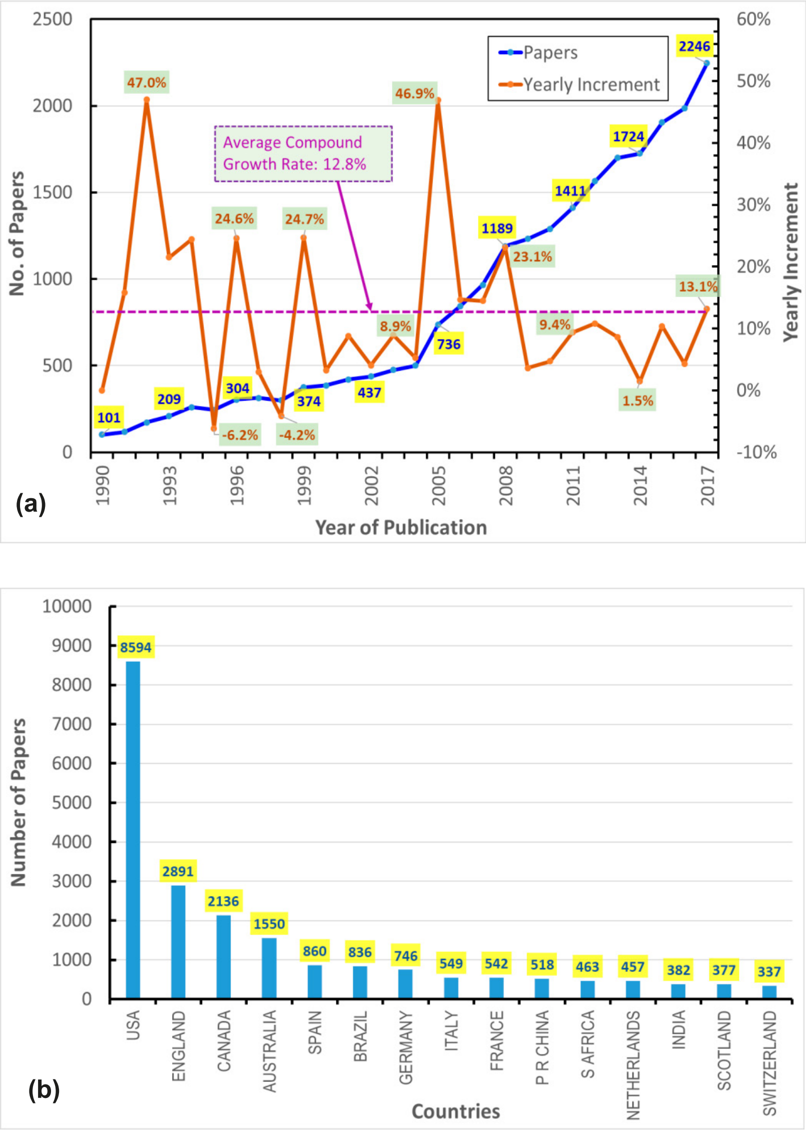


Figure 1 – (a) Yearly publication trends of publications in RIRE-related topics. The primary y-axis represents the number of papers per year, while the secondary y-axis shows the yearly increment. (b) Publication count on RIRE topics from 1990-2018 by country.

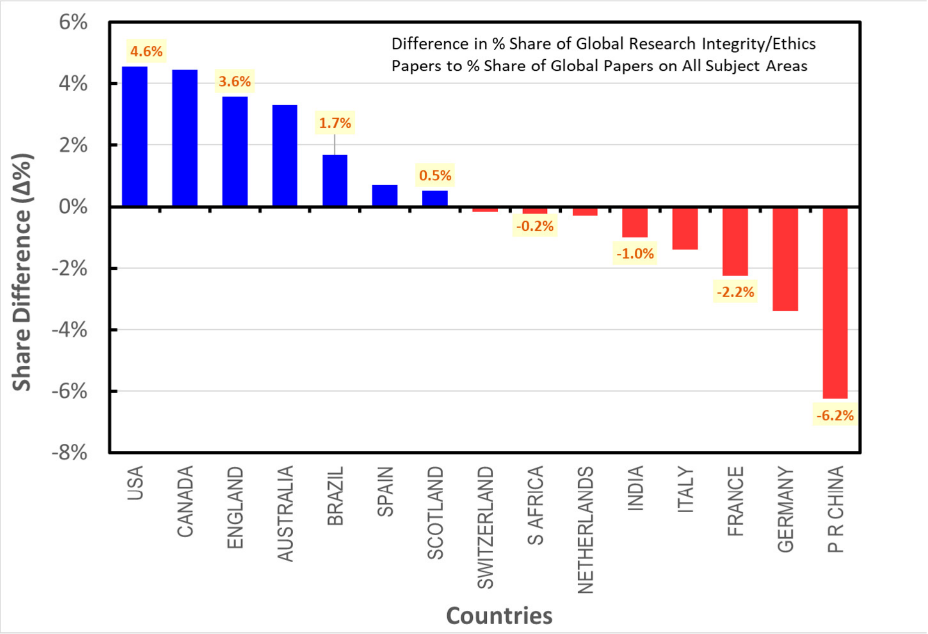


Figure 2. Share difference in global publications on RIRE against All Research Areas for top countries between 1990-2018.

A comparative analysis is further conducted on the two leading publishers globally, USA and China, along with a comparison between emerging countries (China and India). Figure 3(a) shows the global shares of RIRE publications compared with global shares from all research areas for both USA and China. Compared to the world output, USA exhibited declines in both categories. Meanwhile, publication shares of China rose from less than 1% in 1990 to 17.9% in 2018, ranking second in research publication output in WoS. However, although China has grown its research capacity dramatically in the past three decades, its participation in RIRE-related work is comparatively lower (3.5%) and not commensurate with its overall growth (Ataie-Ashtiani 2017, 2018; Qiu 2010; Yang 2013).

The current ratio of Chinese publications on RIRE relative to the country's total number of publications in 2018, marked at 0.016%, is equal to that of the USA in 1990. The Chinese government has taken firm steps to respond to the problems of RIRE deficiency (Zeng & Resnik, 2010; Qiu, 2015; Yi, Standaert, & Nemery et al., 2017; Cyranoski, 2018), such as developing ethics policies and establishing oversight committees. An extensive punishment system was announced that could have significant consequences for offenders — far beyond their academic careers (Cyranoski, 2018). Offending researchers could face restrictions on jobs, loans, and business opportunities. With these measures, improvement on RIRE is anticipated. However, most importantly, research and education in RIRE should be enhanced (Zeng, & Resnik, 2010), so that a healthy research community with high standard research ethics can be established.

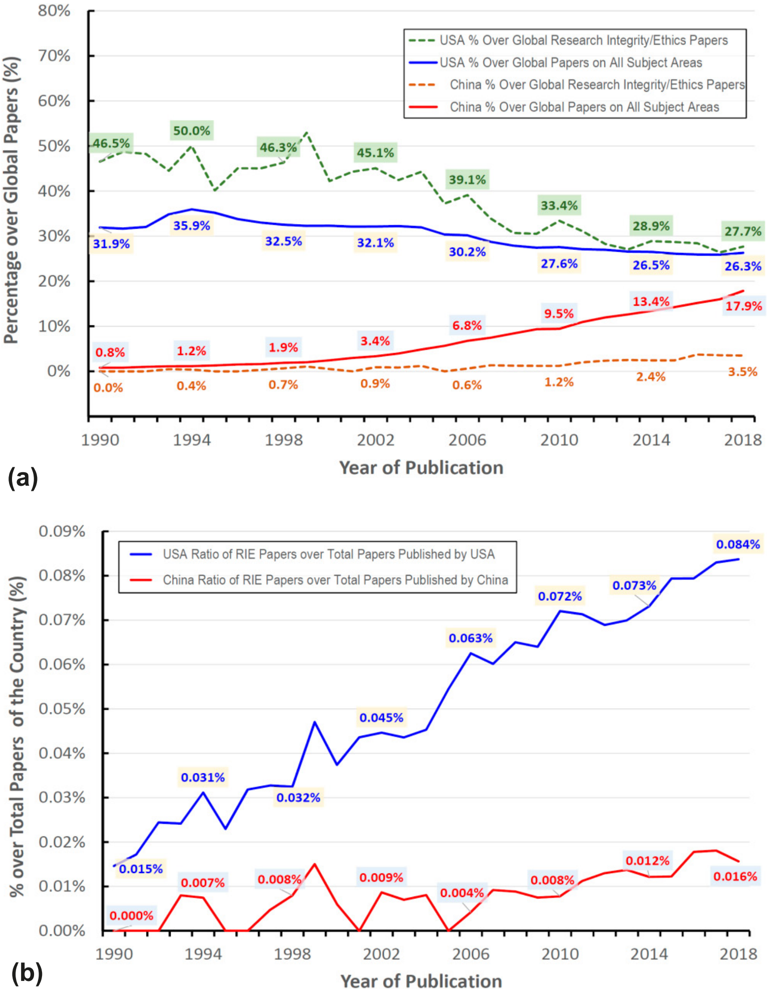


Figure 3 – (a) Proportion of US and Chinese publications globally on RIRE (dashed) compared with proportions in all research areas of global publication (solid). (b) Proportion of US and Chinese publications against total publications in all research areas.

Figures 4(a) and (b) show a comparison of global shares in RIRE publications relative to all research areas held by China and India. While India and China have comparable publications shares on RIRE, China has a far greater growth rate on publications in all research areas (Bhattacharya, Shilpa, & Kaul, 2015).

Based on the global keyword co-occurrence mapping, some of the key RIRE topics gravitate around research ethics, bioethics, and integrity (Figure 5). Research ethics addresses the protocols and policies for adherence in conducting scientific investigation. Many pertain to clinical research methodology matters, with terms such as 'institutional review board', 'informed consent', 'privacy', and 'ethics committees'. Bioethics represents a specialized case addressing ethics in biomedical and medical research, such as social values in medical ethics (e.g. decision-making, dignity, euthanasia, death, quality of life) or the intersection of research with biotechnology (e.g. cloning, stem cells). The dominance of medical-related ethics is further observed based on journal frequency, where *BMJ Open*, *J Med Ethics*, *AM J Bioethics*, and *Bioethics* are some of the most common journals amongst the publications. Another cluster relates to matters of research integrity, prominently on ‘plagiarism’ and ‘scientific misconduct’. Plagiarism, taking credit for the intellectual work of another, is seen in both education and research spheres. Works on the former include studies related to student academic integrity, such as cheating. Scientific misconduct refers to authorship matters with keywords such as publication transparency, publication ethics, and conflict of interest. This forms a comparatively smaller subset, but has been a topic of growing interest in the last 20 years (Aubert Bonn and Pinxten 2019).

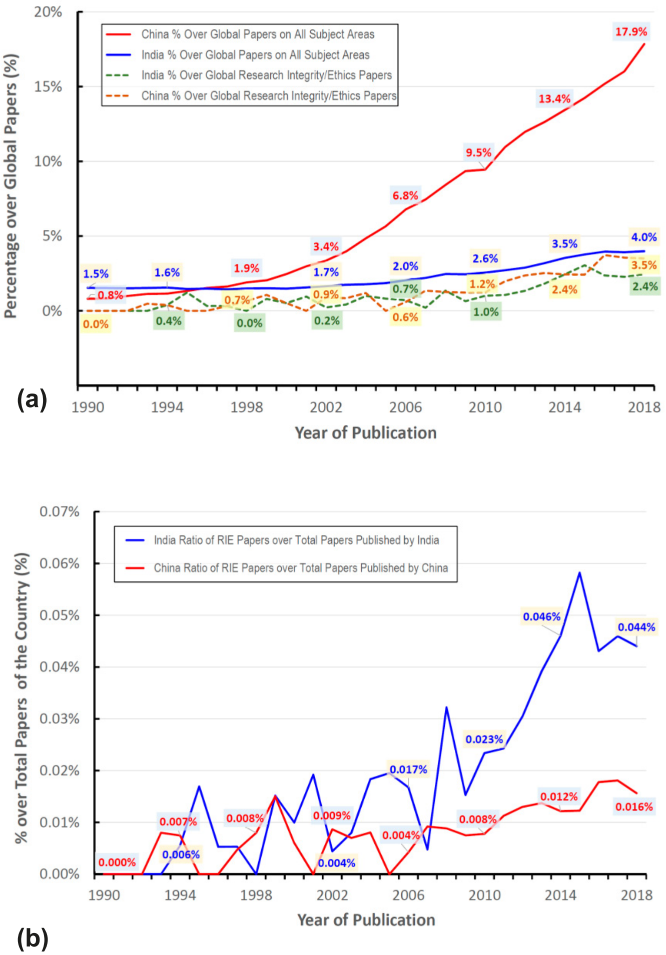


Figure 4 – (a) Proportion of publications from China and India globally on RIRE (dashed) compared with proportions in all research areas of global publication (solid). (b) Proportion of publications from China and India against total publications in all research areas.

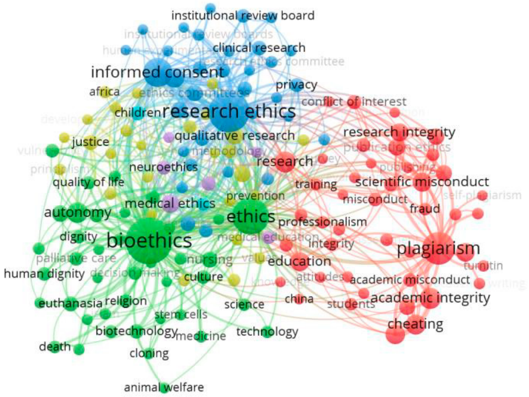


Figure 5 – Author keyword co-occurrence network mapping, showing the three main themes in RIRE research between 1990-2018: research ethics, bioethics, and integrity.

Country collaboration patterns based on co-authorship networks show that countries leading in publications are engaged in international collaboration and form a central part of a cluster (Figure 6). USA exhibits the greatest centrality compared to the other top 10 nations. The node proximity of the collaboration map suggests that western European countries work more closely, likely facilitated by geographic proximity. Meanwhile, another cluster is identified to consist mostly of Spanish-speaking countries (e.g. Spain, Mexico, Colombia, Argentina, Chile), linked to Brazil. China and India most closely collaborate with Southeast Asian, African, and Middle Eastern nations. Topical inspection of network maps by each country reveals variations in the RIRE focus. Publications in wealthier nations such as USA, England, and Canada focus publications more on academics (Dwan et al. 2008; McCabe et al. 2001, 2006), social sciences (Minkler 2004), medical, and biosciences (Begley and Ioannidis 2015; Dickersin and Rennie 2003; Hopewell et al. 2009; Wolf et al. 2012). Those of middle-income countries such as Brazil and South Africa center around social science and healthcare matters (Cabral et al. 2006; de Castilho and Kalil 2005; Diniz et al. 1999), such as ethics related to healthcare policies or protocols (Benatar and Singer 2000, 2010; Shapiro and Benatar 2005). Meanwhile, publications from China are associated with higher education (Gray and Jordan 2012) and research misconduct matters (Lin 2013; Macfarlane et al. 2014).

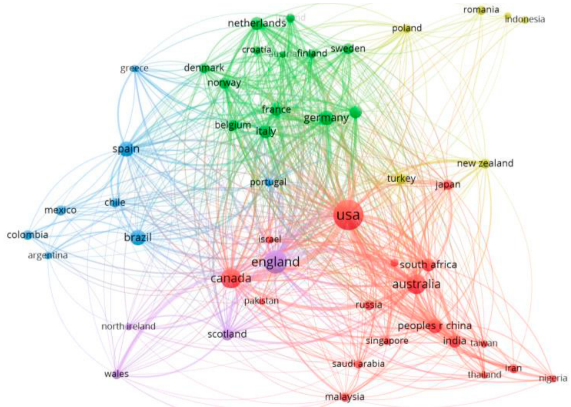


Figure 6. Country co-authorship network map between 1990-2018.

Figure 7(a) shows yearly trends of publication output which are international collaborations by absolute count (primary axis), and by proportion of all publications (secondary axis). International collaboration grew from around 2% of total publications in 1990 to 23% of total publications in 2018. Most collaborations arise from Western nations, with the top 10 consisting of pair combinations between the USA, Canada, England, Australia, England, Scotland, Wales, France, Netherlands, and Germany (Figure S1). In terms of authorship, single-authored papers dropped from 70.4% in 1990-2004 to 36.3% in 2015-2018, while papers with 4 authors increased 3-fold, and papers with 8 and 10 authors increased by one order of magnitude. Single-institutional papers dropped from 82.1% in 1990-2004 to 45.1% in 2015-2018, while papers with 4 institutions increased 5-fold, and papers with 8 institutions have increased by one order of magnitude. Single-country papers dropped form 95.4% in 1990-2004 to 81.1% in 2015-2018, while papers with 2 countries have almost increased 4-fold. In all cases, RIRE publications are becoming more collaborative institutionally, nationally, and internationally (Figure 7(b)-(d)).
We note that there are some limitations to the study. The skew towards the natural sciences and engineering within the database may result in the analysis to reflect these fields more dominantly, while underrepresenting others such as the social sciences and humanities. This is likely to have contributed to the stronger representation of medical ethics-related works, which is part of the database curation.

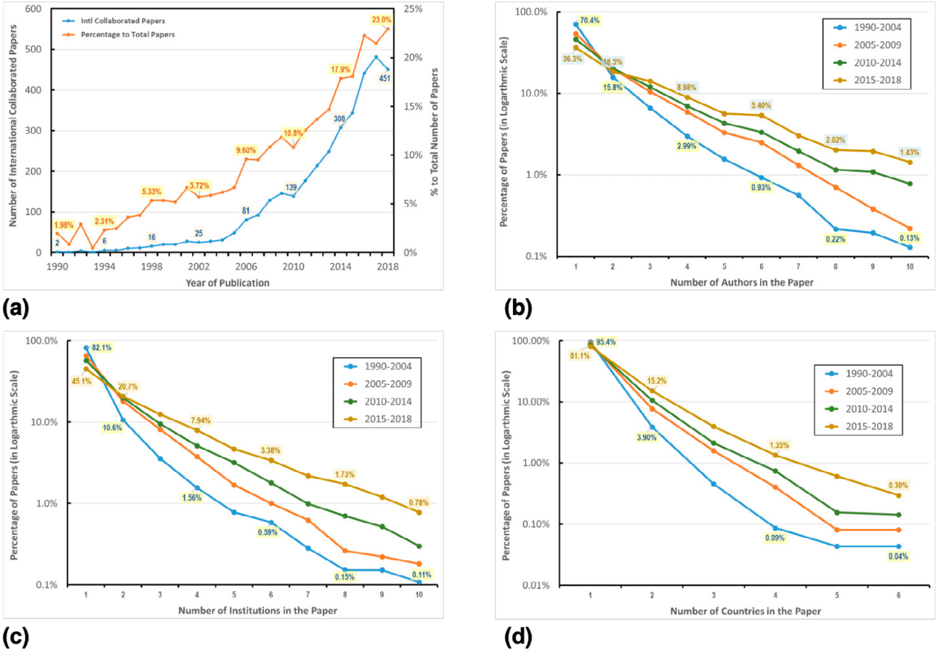


Figure 7. (a) International collaborations in RIRE publications between 1990-2018 and trends at the (b) author- (c) institutional- and (d) country-level.

Conclusions

In this study, bibliometric analysis was used to capture and characterize the macroscopic trends within the broader RIRE research community. The results indicate that while RIRE productivity has continued to grow between 1990-2018, there are variations between countries on factors such as topics of publication and rate of growth. A number of key areas are found within RIRE research, including research integrity, bioethics, scientific misconduct, and plagiarism, which exhibit some geographic dependence, with countries of greater centrality exhibiting a broader scope of topic coverage. Top contributors of RIRE research and collaboration partners are generally dominated by wealthier nations such as the United States, England, Canada, Australia, Germany, and France, with typical collaboration pairings amongst fellow top publishers such as US-Canada and US-England. Although emerging countries have a growing presence in global research participation on all subjects, and in some cases outpacing more developed and counterparts, RIRE-related publications do not follow the subject share proportion growth rates in the same manner. For example, China exhibited meteoric growth in research and innovation, becoming the second leading country in science and technology, with publications from China accounting for a 17.9% share of global papers on all subjects. However, its RIRE growth has not been accompanied in equal measure, holding a smaller global subject share of 3.5%.

Based on current trends, we expect there will be a sustained interest and discourse in RIRE topics as research integrity and research ethics are universal concerns (L. Bouter 2020) for several reasons. First, the credibility and research reputation of a university is intimately tied to its ability produce quality research with good accountability (Hudson 2008). There is added visibility from publicly available databases which monitor research activities, such as Retraction Watch. Currently, institutions (Aubert Bonn et al. 2017) and countries have substantially different norms and standards governing scientific work (Leshner, & Turekian, 2009) and international collaboration. We found that international collaboration on RIRE increased from around 2% of the total publications in 1990 to 23% of total publications in 2018. The relative decrease in single author papers, single institution papers, and single country papers, and the increase of two or multiple author/institution/country papers over time, indicates that joint papers in RIRE are involving more countries, institution, and authors. As more institutions adopt initiatives such as formal structured training (DuBois et al. 2008; Ferguson et al. 2007; Satalkar and Shaw 2019; Sponholz 2000), there may be more opportunities for further collaboration and communication towards harmonizing standards to promote more uniform global research practices (Frankel, Leshner, & Yang, 2016). Secondly, the scope of topics is expected to change depending on factors such as relevance to the publishing country or funding agency, and forthcoming applications or technologies. As an example, bioethics is largely affected by the developments within the sector, such as stem cells and cloning. With the integration of artificial intelligence, machine learning, and big data permeating sectors including the medical practice (e.g. patient data records and collection, genomic data and physiologic data), we expect that the future discourse would evolve accordingly, dictated by the prevailing innovations and commercialization.

Acknowledgments

This study is supported by the Ministry of Education Singapore AcRF Tier 1 grant RG141/17.

References

Ana, J., Koehlmoos, T., Smith, R., & Yan, L. L. (2013). Research Misconduct in Low- and Middle-Income Countries. *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.1001315

Ataie-Ashtiani, B. (2017). Chinese and Iranian Scientific Publications: Fast Growth and Poor Ethics. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-016-9766-1

Ataie-Ashtiani, B. (2018). World Map of Scientific Misconduct. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-017-9939-6

Aubert Bonn, N., Godecharle, S., & Dierickx, K. (2017). European Universities' Guidance on Research Integrity and Misconduct: Accessibility, Approaches, and Content. *Journal of Empirical Research on Human Research Ethics*.

Aubert Bonn, N., & Pinxten, W. (2019). A Decade of Empirical Research on Research Integrity: What

Have We (Not) Looked At? *Journal of Empirical Research on Human Research Ethics*. https://doi.org/10.1177/1556264619858534

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*. https://doi.org/10.1161/CIRCRESAHA.114.303819

Benatar, S. R., & Singer, P. A. (2000). A new look at international research ethics. *British Medical Journal*. https://doi.org/10.1136/bmj.321.7264.824

Benatar, S. R., & Singer, P. A. (2010). Responsibilities in international research: A new look revisited. *Journal of Medical Ethics*. https://doi.org/10.1136/jme.2009.032672

Bhattacharya, S., Shilpa, & Kaul, A. (2015). Emerging countries assertion in the global publication landscape of science: a case study of India. *Scientometrics*. https://doi.org/10.1007/s11192-015-1551-4

Bird, S. J. (2006). Research ethics, research integrity and the responsible conduct of research. *Science and Engineering Ethics*, *12*(3), 411–412. https://doi.org/10.1007/s11948-006-0040-9

Bouter, L. (2020). What Research Institutions Can Do to Foster Research Integrity. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00178-5

Bouter, L. M., Tijdink, J., Axelsen, N., Martinson, B. C., & ter Riet, G. (2016). Ranking major and minor research misbehaviors: results from a survey among participants of four World Conferences on Research Integrity. *Research Integrity and Peer Review*. https://doi.org/10.1186/s41073-016-0024-5

Braun, R., Ravn, T., & Frankus, E. (2020). What constitutes expertise in research ethics and integrity? *Research Ethics*. https://doi.org/10.1177/1747016119898402

Cabral, M. M. L., Schindler, H. C., & Abath, F. G. C. (2006). Regulations, conflicts and ethics of medical research in developing countries. *Revista de Saude Publica*. https://doi.org/10.1590/S0034-89102006000300022

Coxon, C. H., Longstaff, C., & Burns, C. (2019). Applying the science of measurement to biology: Why bother? *PLOS Biology*, *17*(6), e3000338. https://doi.org/10.1371/journal.pbio.3000338

de Castilho, E., & Kalil, J. (2005). Ética e pesquisa médica: princípios, diretrizes e regulamentações / Ethics and medical research: principles, guidelines, and regulations. *Rev soc bras med trop*.

Dickersin, K., & Rennie, D. (2003). Registering Clinical Trials. *Journal of the American Medical Association*. https://doi.org/10.1001/jama.290.4.516

Diniz, D., Guilhem, D. B., & Garrafa, V. (1999). Bioethics in Brazil. In *Bioethics*. https://doi.org/10.1111/1467-8519.00152

DuBois, J. M., Dueker, J. M., Anderson, E. E., & Campbell, J. (2008). The development and assessment of an NIH-funded research ethics training program. *Academic Medicine*. https://doi.org/10.1097/ACM.0b013e3181723095

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A. W., Cronin, E., et al. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0003081

Fanelli, D. (2013). Why Growing Retractions Are (Mostly) a Good Sign. *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.1001563

Fanelli, D., Costas, R., & Larivière, V. (2015). Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0127556

Ferguson, K., Masur, S., Olson, L., Ramirez, J., Robyn, E., & Schmaling, K. (2007). Enhancing the culture of research ethics on university campuses. *Journal of Academic Ethics*. https://doi.org/10.1007/s10805-007-9033-9

Geller, G., Boyce, A., Ford, D. E., & Sugarman, J. (2010). Beyond “compliance”: The role of institutional culture in promoting research integrity. *Academic Medicine*. https://doi.org/10.1097/ACM.0b013e3181e5f0e5

Gray, P. W., & Jordan, S. R. (2012). Supervisors and Academic Integrity: Supervisors as Exemplars and Mentors. *Journal of Academic Ethics*. https://doi.org/10.1007/s10805-012-9155-6

Haven, T. L., Bouter, L. M., Smulders, Y. M., & Tijdink, J. K. (2019). Perceived publication pressure in Amsterdam: Survey of all disciplinary fields and academic ranks. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0217931

Hopewell, S., Loudon, K., Clarke, M. J., Oxman, A. D., & Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.MR000006.pub3

Hudson, R. (2008). Research Ethics and Integrity — The Case for a Holistic Approach. *Research Ethics*. https://doi.org/10.1177/174701610800400402

Javed, S. A., & Liu, S. (2018). Predicting the research output/growth of selected countries: application of Even GM (1, 1) and NDGM models. *Scientometrics*, *115*(1), 395–413. https://doi.org/10.1007/s11192-017-2586-5

Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*. https://doi.org/10.1007/s11192-010-0202-z

Lin, S. (2013). Why serious academic fraud occurs in China. *Learned Publishing*. https://doi.org/10.1087/20130105

Lins, L., & Carvalho, F. M. (2014). Scientific Integrity in Brazil. *Journal of Bioethical Inquiry*. https://doi.org/10.1007/s11673-014-9539-y

Macfarlane, B., Zhang, J., & Pun, A. (2014). Academic integrity: a review of the literature. *Studies in Higher Education*. https://doi.org/10.1080/03075079.2012.709495

McCabe, D. L., Butterfield, K. D., & Treviño, L. K. (2006). Academic dishonesty in graduate business programs: Prevalence, causes, and proposed action. *Academy of Management Learning and Education*. https://doi.org/10.5465/AMLE.2006.22697018

McCabe, D. L., Treviño, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: A decade of research. *Ethics and Behavior*. https://doi.org/10.1207/S15327019EB1103_2

Michalek, A. M., Hutson, A. D., Wicher, C. P., & Trump, D. L. (2010). The costs and underappreciated consequences of research misconduct: A case study. *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.1000318

Minkler, M. (2004). Ethical challenges for the “outside” researcher in community-based participatory research. *Health Education and Behavior*. https://doi.org/10.1177/1090198104269566

Moiwo, J. P., & Tao, F. (2013). The changing dynamics in citation index publication position China in a race with the USA for global leadership. *Scientometrics*. https://doi.org/10.1007/s11192-012-0846-y

Okonta, P. I., & Rossouw, T. (2014). Misconduct in research: A descriptive survey of attitudes, perceptions and associated factors in a developing country. *BMC Medical Ethics*. https://doi.org/10.1186/1472-6939-15-25

Powell, J. J. W., Baker, D. P., & Fernandez, F. (2017). *The Century of Science: The Global Triumph of the Research University*. *International Perspectives on Education & Society*.

Qiu, J. (2010). Publish or perish in China. *Nature*. https://doi.org/10.1038/463142a

Resnik, D. B. (2020). What Is Ethics in Research & Why Is It Important? *National Institute of Environmental Health Sciences*. https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm. Accessed 22 April 2021

Satalkar, P., & Shaw, D. (2019). How do researchers acquire and develop notions of research integrity? A qualitative study among biomedical researchers in Switzerland. *BMC Medical Ethics*. https://doi.org/10.1186/s12910-019-0410-x

Schüklenk, U. (2000). Protecting the vulnerable: Testing times for clinical research ethics. *Social Science and Medicine*. https://doi.org/10.1016/S0277-9536(00)00075-7

Shapiro, K., & Benatar, S. R. (2005). HIV prevention research and global inequality: Steps towards improved standards of care. *Journal of Medical Ethics*. https://doi.org/10.1136/jme.2004.008102

Shashnov, S., & Kotsemir, M. (2018). Research landscape of the BRICS countries: current trends in research output, thematic structures of publications, and the relative influence of partners. *Scientometrics*. https://doi.org/10.1007/s11192-018-2883-7

Sponholz, G. (2000). Teaching scientific integrity and research ethics. In *Forensic Science International*. https://doi.org/10.1016/S0379-0738(00)00267-X

Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why Has the Number of Scientific Retractions Increased? *PLoS ONE*. https://doi.org/10.1371/journal.pone.0068397

Titus, S. L., Wells, J. A., & Rhoades, L. J. (2008). Repairing research integrity. *Nature*. https://doi.org/10.1038/453980a

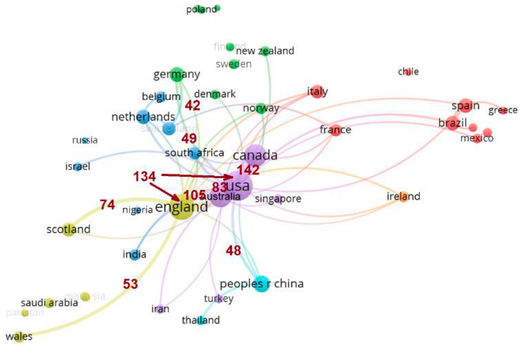
VOSviewer. (2020). Welcome to VOSviewer. *Centre for Science and Technology Studies, Leiden University*.

Wolf, S. M., Crock, B. N., Van Ness, B., Lawrenz, F., Kahn, J. P., Beskow, L. M., et al. (2012). Managing incidental findings and research results in genomic research involving biobanks and archived data sets. *Genetics in Medicine*. https://doi.org/10.1038/gim.2012.23

Yang, W. (2013). Research integrity in China. *Science*. https://doi.org/10.1126/science.1247700

Zhang, L., Powell, J. J. W., & Baker, D. P. (2015). Exponential Growth and the Shifting Global Center of Gravity of Science Production, 1900–2011. *Change: The Magazine of Higher Learning*. https://doi.org/10.1080/00091383.2015.1053777

Supplementary Figures



Top Bilateral Collaborations		Papers	Top Bilateral Collaborations		Papers
USA	Canada	142	England	Germany	42
USA	England	134	England	Ireland	40
England	Australia	105	USA	France	39
USA	Australia	83	USA	Netherlands	36
England	Scotland	74	Germany	Switzerland	35
England	Canada	55	England	Italy	33
England	Wales	53	USA	Switzerland	31
England	France	51	England	Switzerland	31
England	Netherlands	51	Germany	Netherlands	31
Canada	Australia	51	England	Belgium	30
USA	Germany	49	USA	India	28
USA	P R China	48	England	Norway	28
USA	South Africa	43	Italy	France	25

Figure S1. International Bilateral Collaborated Publications on RIRE during 2015-2018.

A Unique Approach to Characterizing Types of Prolific Authors

Richard Klavans[1] and Kevin W. Boyack[2]

[1] *rklavans@mapofscience.com*
SciTech Strategies, Inc., Wayne, PA, (USA)

[2] *kboyack@mapofscience.com*
SciTech Strategies, Inc., Albuquerque, NM, (USA)

Abstract

How do research communities differ in terms of their mix of different types of authors? We introduce a unique approach to characterizing prolific authors in terms of their roles in research communities as reflected by their publication profiles. This approach builds upon Kuhnian concepts of incommensurability (members of one research community do not fully understanding the conceptual framework of another research community) and the more recent literature on types of experts (hedgehogs and foxes). To illustrate these concepts, we use bibliometric data to characterize highly prolific authors as hedgehogs (they know a lot about a single research community), foxes (they know a lot about multiple research communities) and non-conformist (what they know doesn't fit with the way research communities have been defined). Data on the prolific authors who publish in *Scientometrics* and one of the reviews of the first draft of this article are used to illustrate the potential for this approach.

Introduction

The ability to use very large document collections to characterize tens of thousands of Kuhnian research communities has opened up new avenues of inquiry. For instance, we can now explore the sociological understanding of how these communities differ from each other and the corresponding insights that can arise from this understanding.

A prime example of this paradigmatic framework is the recent work at NIH where a text-based method (word2vec) was used to cluster NIH proposals into document clusters (Hoppe et al., 2019). Each document cluster, since it uses a different vector of words, is considered to be a different conceptual framework. Once these frameworks were identified, it was then discovered that African-American/black researchers tended to propose research on topics with lower award rates. This analysis provided unique insights into addressing racial bias in the funding of US medical research. Racial bias was associated with underlying biases in conceptual frameworks and can be addressed by a better understanding of these implicit psychological biases (Livingston, 2021) and redirecting funding to these research communities.

Another example of this paradigmatic framework is the recent work at the Center for Security and Emerging Technology (CSET), a think tank that was started in 2019 by Jason Matheny (former Director of IARPA and currently serving in the Biden Administration). One of their recent publications (Rahkovsky, Toney, Boyack, Klavans, & Murdick, 2021) identified 127,000 Kuhnian research communities using 1.4 billion citation links among 105 million documents from Dimensions, Microsoft Academic Graph, Web of Science and the Chinese National Knowledge Infrastructure. This analysis resulted in specific forecasts about the research communities in Artificial Intelligence that are expected to have exceptional growth in the next three years. These nominated research communities represent an emerging threat or opportunity. These forecasts are correspondingly tested using techniques that are commonly used to evaluate the accuracy of weather forecasts.

The purpose of this study is to introduce a unique author-level measure that can be used in studies such as the two mentioned above. We rely on bibliometric information about the research communities that an author publishes in order to classify authors as hedgehogs, foxes and non-conformists. The following two sections describes the related literature and the theoretical framework for the proposed measures. We then proceed with a description of a

Kuhnian model (100,000 research communities derived from a direct citation analysis of the Scopus database) and the classification of prolific researchers who have published in journal *Scientometrics* in recent years. One of the reviews of our first draft is also used to illustrate how this indicator might be useful to understanding how research communities are evolving. The final section describes the limitations and implications of this study.

Empirical Precedent

There is empirical precedent for studying types of individuals that populate a research community. For instance, Robert and Fusfeld (1980) identified critical roles (i.e., "people functions") in a team of researchers in a research lab. Studies that describe author contributions to an article (Sauermann & Haeussler, 2017) described the different hats worn by (or roles of) an individual scientist such as obtaining funding, planning, running experiments and writing. Journal editors are often thought of as having the role of gatekeepers (Crane, 1967). The roles of scientists in interfacing with society have also been explored (Spruijt et al., 2014).

There is also an extensive bibliometric literature on how one might characterize different authors using different bibliometric indicators. They can be characterized by how prolific they are and whether their activities are focused or dispersed (Abramo, D'Angelo, & Di Costa, 2017, 2019). They can be characterized by how quickly they have reached a certain level of productivity (Abramo, D'Angelo, & Di Costa, 2018; Costas, Van Leeuwen, & Bordons, 2010) and the degree to which they are involved in team science (Wuchty, Jones, & Uzzi, 2007). One recent review listed over 100 bibliometric-based indicators that have been used to characterize individual authors (Wildgaard, Schneider, & Larsen, 2014).

There are, however, no studies that use the bibliometric data on prolific authors to characterize the role they might play as an author, a leader or perhaps even as a reviewer in a specific research community, and the corresponding number of research communities where the author would be considered an expert in. The following section describes the theoretical basis for creating such an indicator.

Theory

This study builds on Kuhn's (1970) concept of a research community and Tetlow's (2006) characterization of experts as hedgehogs and foxes.

Kuhn's theory, as originally proposed, was based on the possibility that the norms of good science are social and can be detected by examining the communication patterns between researchers. Although he viewed research communities as groups of authors rather than groups of documents, he was well aware of Garfield's new (at that time) citation database and noted that direct citations could be used to detect research communities due to the fact that they could be considered as evidence of communication links between researchers. Much of the early work on science mapping – clustering documents using citation links – was based on this theoretical root because it was viewed as a proxy for communication.

We are returning to this theoretical root because it appears that most current document clustering exercises (including some of ours) have been aimed at identifying scientific 'topics' without explicitly defining what a topic is and without reference to any underlying theory (Held, Laudel, & Gläser, 2021). Further, we seem to have lost sight of the researchers themselves, that they view themselves as belonging to communities, that each researcher may have a different perception of a given community and its organizing principle, and that the inside and outside views of a community may be different.

We do not advocate the clustering of authors to identify Kuhnian research communities. If one clusters authors, the organizing basis of the community (i.e., the focal point, theory, method, topic, etc.) must be determined from other features, most commonly from documents and their metadata. Even then it may be difficult to identify the focal point of a cluster of authors given that some authors are prolific and can be identified as members of multiple communities. Clustering of documents is more straight-forward in that the focal point of a cluster of documents can be identified more readily and with less ambiguity. Citation links provide a ready source of information with which to cluster, and this brings us back full circle to Kuhn. The difference may be, however, that the current generation of researchers start with methods and gives lip service to theory while the early pioneers of our field put theory first and then proceeded to methods.[1]

We correspondingly use clusters of documents based on direct citation to identify Kuhnian research communities, while explicitly remembering that it is the relationship between citations and communication patterns that is the basis for claiming that document clusters derived from citation patterns may represent Kuhnian research communities. Direct citations are evidence of communications between members of a research community but not because the citing author is communicating with the cited author. The communication link is between the author and the reviewers of the article. Reviewers are usually selected because they have some knowledge about the thing that the author is talking about. It is not uncommon for reviewers to point out gaps in the reference lists. And it is not uncommon for authors to realize that a highly critical review is, in many cases, a signal that the author did not effectively communicate the unique perspective of the research community that the document is aimed at. It is this dialogue between an author and his peer reviewers that is central to the process of community evolution. But all we can directly observe are the negotiated set of cited references from this dialogue. Overall, it is the relationship between citations and communication patterns that is the basis for claiming that document clusters, derived from citation patterns, are a reasonable way to represent Kuhnian research communities.

We are also building on Tetlow's [2006] study of types of experts in order to better characterize the types of reviewers that exist in a research community. Tetlow, in an extensive study of the accuracy of expert forecasts within the political domain, noted that some experts 'knew a lot about one thing' (hedgehogs) while others 'knew a lot about many things' (foxes). This concept is applied to prolific authors. An author that focuses their effort in a single research community could be considered a hedgehog when they are asked to review a paper. An author that has an extensive publication record in two or more research communities could be considered a fox (than know a lot about many things). And perhaps most importantly, those highly prolific authors who don't focus their publications in any research community represent a third type of reviewer (not anticipated by Tetlow). These experts aren't conforming to the social norms of research and are therefore considered non-conforming.

In summary, Tetlow's concept of hedgehogs and foxes is used to characterize the types of reviewers that exist in a research community. The link between these two theoretical perspectives is used to better understand the composition of research communities (percentage of hedgehogs, foxes and non-conformists) and how differences in this composition is related to broader issues about the structure and evolution of research.

Data and Methods

This study is based on Scopus data and is enabled by 1) our existing detailed model of research communities and 2) Scopus author profiles and the assignment of author-ids to individual papers (Baas, Schotten, Plume, Côté, & Karimi, 2020).

[1] We acknowledge Jochen Gläser for discussions in which this became much more clear. After reviewing many of the studies of document clustering, he has concluded that advocates of different document clustering algorithms have lost sight of what their theoretical assumptions are. We agree with his assessment, not because there isn't any theory that is applicable to the document clustering problem, but rather because those that do this work describe their theoretical framework in a perfunctory manner. As such, the theory section of this paper is intended to redirect our thinking toward theory in light of decades of neglect.

Our model consists of 46,100,988 papers indexed by Scopus (1996-2019) that have references or are cited and thus can be linked into a cluster of papers via citation patterns. The model also contains over 44.5 million additional documents cited at least twice by the indexed papers. These so-called non-source documents are those for which we do not have references and include documents not indexed by Scopus along with some documents from before 1996 that do appear in Scopus. Clustering of documents was done using our established process (Klavans & Boyack, 2017) with the Leiden algorithm (Traag, Waltman, & Van Eck, 2019). The input consisted of over 1.13 billion direct citation links between the 90+ million documents, resulting in a model containing 104,638 clusters (representing research communities), each of which contained at least 75 documents.

As subject for our case study we selected the 14 research communities where *Scientometrics* is the dominant journal. From the perspective of *Scientometrics*, these 14 research communities represent the areas of research where the journal has the highest relative influence. Characteristics of these research communities, including the number of papers, number of unique authors, and least-squares growth rate are given in Table 1. The publication output of these communities varies widely. From 2015 to 2019, over 500 papers per year were published on issues around research evaluation (RC #365) while less than 12 papers per year were published on analyzing funding acknowledgements. (RC #39409).

Table 1. Characteristics of 14 research communities where *Scientometrics* is the lead journal. Time period 2015-2019. *%Grow* is the growth rate.

RC	*Phrases (description)*	*#Doc*	*#Auth*	*%Grow*
365	Academic rank / citation counts/ impact factor / Hirsch index	2599	5023	-2.5%
9183	Alternative metrics / citation counts / Mendeley reader count	1034	1917	10.2%
5104	Co-authorship network / scientific collaboration	939	2007	2.9%
4581	Peer review / peer review process / open peer review	838	1773	-8.3%
6344	Co-word analysis / science mapping / intellectual structure	782	1817	6.5%
6864	Interdisciplinary research / team science	708	2081	6.6%
15058	Top 100 cited articles in …	581	1757	6.5%
15746	Norwegian model / national databases / scholarly books	350	460	6.6%
13588	(Bibliometric overviews of single journals)	234	454	18.5%
9942	(Bibliometrics in economics)	217	356	-1.7%
12107	Web impact factor / hyperlink network / webometric analysis	155	294	-7.5%
37671	Title length / research titles / conceptual diversity	66	144	-0.2%
81246	Reference publication year spectroscopy	62	109	23.7%
39409	Funding acknowledgements / funding ratios / funding data	57	120	25.6%

We note that three of the research communities in Table 1 may not be well known to most researchers in the scientometrics community. RC #15058 is typified by papers with titles such as "Top 100 cited articles in <a branch of medicine>" that are scattered throughout the medical literature. RC #13588 contains papers that do a cursory analysis of the output of a single journal while RC #9942 is focused on bibliometrics as applied to economics journals and researchers. In each of these research communities the paper counts per journal are very low (typically single digits) and they appear in our list simply because *Scientometrics* has the largest number of papers even though that number is small.

The number and frequency of papers published by different authors varies widely. Ioannidis et al. (2014) showed that authors who published continuously over a 16 year period, while a very small percentage (<1%) of the author population, account for over 40% of papers, garner far more citations per paper than less prolific authors and are therefore disproportionally influential. On the lower end, Price & Gürsey (1975) pointed out that a large fraction of authors only publish one paper; in their sample 56% of authors only published one paper in a seven year period. We found that 19.7% (3246/16483) of the authors accounted for in Table 1 only published one paper during the 2015-2019 time period while 37.3% (6156/16483) published at least 12 articles over the same period. The following analysis focuses specifically on these prolific authors.

There are no set criteria for determining what constitutes a prolific author and what level of publication in a research community represents 'knowing about' that research community. We have arbitrarily set a threshold of 12 publications over the four-year period from 2015-2018 (or three papers per year on average) to identify prolific authors. In addition, we specify that an author must publish at least four publications in a research community over that four-year period as evidence that she 'knows about that research community'. We believe that these publication rates are sufficient to classify an individual as a hedgehog, a fox, or neither. Given that our case example focuses on *Scientometrics*, we limit our analysis to those authors who have at least 50% of their publications during the time period in the 14 research communities listed in Table 1.

Prolific authors in scientometrics were classified into one of four types as follows:

- *Hedgehog* – at least four papers in only one research community
- *Fox2* – at least four papers in exactly two research communities
- *Fox3+* – at least four papers in three or more research communities
- *Nonconformist* – prolific, but with papers spread out among research communities such that the threshold of four papers in any single community is not met.

We consider hedgehogs and foxes to be the core authors in a community.

Results

Of the 16,483 unique authors that published in our set of 14 research communities we find 93 who meet the criteria mentioned above. Table 2 shows the distribution of author types as a function of the number of articles. Authors with at least three papers but less than four papers per year were most likely to be *hedgehogs*. Authors with at least four papers per year are very unlikely to be *nonconformists* – only one out of 45 authors in our set with at least 16 papers did not have at least four papers in any single research community. Authors with at least five papers per year are more likely to be *foxes* than *hedgehogs*. However, we do see some *hedgehogs* with large numbers of papers. Examples of each author type will be given later.

Table 2. Author type distribution as a function of production.

#Articles (2015-2018)	*#Nonconform*	*#Hedgehog*	*#Fox2*	*#Fox3+*
12-15 (3+ per year)	5	38	5	0
16-19 (4+ per year)	1	9	8	1
20-27 (5,6+ per year)	0	2	6	2
28-39 (7,8,9+ per year)	0	1	2	3
40+ (10+ per year)	0	1	1	8

The distribution of author types by research community is shown in Table 3. In general, larger communities have larger numbers of these prolific authors, with a mix of *foxes* and *hedgehogs*, as might be expected. However, there are some communities that differ from this norm. For example. RC #4581 on peer review has only three *hedgehogs* and no *foxes*, a very small set of core authors for a community with nearly 200 papers per year. RC #13588 on journal studies has no core authors. RC #81246 on research publication year spectroscopy has three *foxes*, and thus has a relatively large and productive set of core authors whose influence might help this community to continue to grow at a high rate.

Table 3. Author type distribution by research community.

RC	*Short label*	*#Doc*	*#Hedgehog*	*#Fox2*	*#Fox3+*
365	Evaluation	2599	18	8	14
9183	Altmetrics	1034	9	8	6
5104	Collaboration	939	3	5	8
4581	Peer review	838	3		
6344	Science mapping	782	3	6	6
6864	Interdisciplinarity	708	1	1	2
15058	Top cited articles	581	3		1
15746	Norwegian model	350	8	5	1
13588	Journal studies	234			
9942	Economics	217	1	1	
12107	Webometrics	155			2
37671	Title length	66		1	
81246	RPYS	62			3
39409	Funding data	57	1		

It is also interesting to look at specific authors and their personal research portfolios. The three most productive authors from each author group are listed in Table 4 along with the research communities in which they have at least four papers. Total papers, numbers of papers in the 14 *Scientometrics* research communities (#Doc14) and the largest number of papers by the author in any community are also listed.

Ronald Rousseau is a prototypical example of a *hedgehog*. He is very productive with 47 papers over the four-year period, 30 of which are in the seven of the 14 Scientometrics communities. His primary area of focus is RC #365 with 21 papers, followed by RC #6344 with three papers. His remaining 23 papers are dispersed among 19 other research communities. Our primary example of a *Fox2* is Felix Moya-Anegón. Of his 28 papers, 24 are in the Scientometrics communities with 16 in RC #365 and 6 in RC #6344. Our third example of a *Fox2*, Stefanie Haustein, is notable in that her second community, RC #3406, is outside of our set of 14 communities. This community is focused on open access, which is certainly highly relevant to the scientometrics community even though *Scientometrics* is not the leading publication venue in the topic.

Lutz Bornmann and Mike Thelwall are two examples of a *Fox3+*. Each has 121 papers over the four-year period and each has at least four papers in six different research communities. Both authors are highly active in one community that is outside our set, RC #36708 (innovation policy) for Bornmann and RC #232 (sentiment analysis) for Thelwall. Those familiar with the works of these authors will recognize their output in these topics. Thelwall's output is more dispersed than Bornmann's in that the former has 36 publications outside our set of communities while the latter has 15.

Table 4. Top three most productive authors from each author type. Research communities outside the list from Table 1 are shown in italic font.

Name	*Type*	*RC#*	*#Doc*	*#Doc14*	*#Large*
Rousseau, R.	Hedgehog	365	47	30	21
Cobo, M.J.	Hedgehog	6344	34	18	11
Sivertsen, G.	Hedgehog	15746	21	19	16
Moya-Anegón, F.	Fox2	365, 6344	28	24	16
Kousha, K.	Fox2	9183, 15746	28	21	9
Haustein, S.	Fox2	9183, *3406*	27	20	16
Bornmann, L.	Fox3+	365, 9183, 81246, 6344, 5104, *36708*	121	106	56
Thelwall, M.	Fox3+	9183, 365, 15746, *232*, 5104, 12107	121	85	35
Leydesdorff, L.	Fox3+	6344, 365, 81246, *630*, 5104, *41710*	74	49	19
Rafols, I.	Nonconf		17	11	3
Soheili, F.	Nonconf		15	8	3
Torres-Salinas, D.	Nonconf		14	8	3

Our primary example of a *nonconformist* is Ismael Rafols. Eleven of his 17 publications were in six of our 14 research communities, with three each in RC #365 and RC #15746. His six remaining publications were in six different research communities dealing with topic modeling, technology foresight, patent analysis and translational science. While there were only six *nonconformists* in our sample, these people may be extremely important because they may represent a fundamentally different 'logic' for deciding which research communities to participate in. 87 of our 93 core authors are focusing their work within one or more research communities. On average, *hedgehogs* have 52% of their papers in their dominant community, *foxes* have 40% of their papers in their dominant community, while *nonconformists* have less than 25% of their papers in a dominant community. *Hedgehogs* and *foxes* thus seem to choose topics in a way that is consistent with the communication patterns assumed by Kuhn to be at the foundation of a research community. *Nonconformists* seem to have broader interests that may indicate a different 'logic' about topic choice.

Discussion

We have significantly revised this paper to illustrate how the proposed indicators relate to the two comments we received by reviewers. The first review we received stated that "the results of the selected example were somewhat debatable" and "thought provoking". This review is non-communicative (no specific comments, suggestions or criticism) and not actionable (the review does not require any revisions on our part).

The second review, however, has had a significant impact on our revisions and is an excellent example of the types of issues we wanted to communicate in this paper. The second reviewer starts with a summary of what we intended- but not what we said in our initial abstract and introduction. This reviewer realized that we were interested in a 'research community' approach. Our abstract and introduction did not stress this fundamental point.

A detailed review followed that mentioned weakness that "could be easily revised in a new version". It started with specifics about articles on individual-level bibliometric indicators and establishing of thresholds. This was followed by a more general observation that the examples we provided are 'illustrative' and ended with the following statement:

> "*The theoretical propositions of the introduction do not come back in the discussion. This is unfortunate because this is one of the strongest points of the manuscript. I would recommend extending the discussion on how more individual-level and*

> *community-based approaches could have more relevance in the light of Kuhn's theory, as well as other related theories.*"

This statement prompted a major revision of the paper. Perhaps more importantly, the comments from the second review may help illustrate some of the issues we are trying to address. Based on the comments we assume that the second reviewer is well versed in Kuhnian theory, individual-level indicators and science mapping, and thus seems to have skills that are found in two research communities: RC#365 (evaluation and bibliometric indicators) and RC#6344 (science mapping, topics, communities). We thus posit that the second reviewer is a fox with a publication record in these two communities and was well qualified to review this work.

We agree with the comments of both reviewers: the results of this study are illustrative rather than conclusive. Our thresholds were based on our experience and our perceptions of norms in science and were not based on theory or data. Thus, we do not argue that the thresholds are correct. Instead, we are simply using an initial set of thresholds to illustrate that there are different authorship profiles within research communities and that, once defined, the mix between these different types of authors (hedgehogs, foxes and non-conformists in this paper) can be determined. It is unclear to us if theory would suggest any particular threshold. Nevertheless, additional research can certainly be undertaken to determine the sensitivity of such results to different thresholds and to quantify the perception of researchers as to what level of publication is considered prolific.

The mix between hedgehogs and foxes in these two research communities (Table 3) may provide insights into how these research communities might evolve. The ratio is 18:22 for RC:365 and 3:12 for RC:6344. Or stated differently, RC:6344 has a far greater percentage of prolific authors that have expertise outside of RC:6344. A research community with a higher proportion of foxes may have more porous boundaries and thus might undergo more structural change in the future.

The use of altmetrics (RC #9183) could help to improve the classification of prolific authors into different types deals with communication outside the traditional realm of citations. RC #5104 deals with collaboration and co-authorship, and it would be interesting to understand if *hedgehogs*, *foxes* and *nonconformists* have different collaboration patterns, and if so, if those different patterns lead to different types of impact (e.g., academic vs. economic). The relationship between author type and transformative research or interdisciplinarity (RC #6864) would also be interesting to funders and other institutions.

Limitations and Future Directions

The analysis presented here is preliminary and is subject to several limitations. First, as mentioned above, we have relied only upon our own opinions as to the reasonableness of the classification. Others may have different opinions, and those could be equally valid to ours. Thus, as mentioned, we plan to seek other opinions. Second, the results are reliant upon the classification system that was used. A less granular set of research communities would obviously increase the number of *hedgehogs* and reduce the number of *foxes* if the same thresholds were used. However, given that we are seeking to identify differences, we favor the use of a large number of highly differentiated research communities. Changes to thresholds would also make a difference in how authors are classified. Fractions of papers by research community could be used rather than number of papers and this would likely reduce the need to split the *fox* type into two types.

Third, this analysis was performed for the single, relatively small research field associated with the journal *Scientometrics*. This area has a relatively low number of authors per paper; certainly the large teams that often author papers in biomedicine and physics are not present in this field.

Fields with larger teams may have different characteristics. In any case, the results here cannot yet be generalized to other fields. This is something we plan to investigate in the future. In addition, author order was not accounted for in this study, and may make a difference in the roles played by authors in communities. Despite these limitations, we find the results to be interesting enough to continue investigation.

In the future we plan to expand this work to identify author types in multiple ways. For instance, we plan to explore the relationship between full and fractional publication counts and how that relates to who is a prolific author. We also hope to explore whether particular research communities exhibit individualistic or collective behavior and the age of researchers in a community to identify so-called human resource profiles at the community level. If successful, these efforts will give us a new way to understand differences in research communities from a person-oriented perspective.

References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2017). Specialization versus diversification in research activities: The extent, intensity and relatedness of field diversification by individual scientists. *Scientometrics, 112*(3), 1403-1418.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2018). The effects of gender, age and academic rank on research diversification. *Scientometrics, 114*(2), 373-387.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2019). Diversification versus specialization in scientific research: Which strategy pays off? *Technovation, 82*(51-57).

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies, 1*(1), 377-386.

Costas, R., Van Leeuwen, T., & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology, 61*(8), 1564-1581.

Crane, D. (1967). The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist, 2*(4), 195-201.

Held, M., Laudel, G., & Gläser, J. (2021). Challenges to the validity of topic reconstruction. *Scientometrics, accepted*.

Hoppe, T. A., Litovitz, A., Willis, K. A., Meseroll, R. A., Perkins, M. J., Hutchins, B. I., . . . Santangelo, G. M. (2019). Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. *Science Advances, 5*, eaaw7238.

Ioannidis, J. P. A., Boyack, K. W., & Klavans, R. (2014). Estimates of the continuously publishing core in the scientific workforce. *PLoS ONE, 9*(7), e101698.

Klavans, R., & Boyack, K. W. (2017). Research portfolio analysis and topic prominence. *Journal of Informetrics, 11*(4), 1158-1174. doi:10.1016/j.joi.2017.10.002

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions* (2nd ed.). Chicago: University of Chicago Press.

Livingston, R. (2021). *The conversation: How seeking and speaking truth about racism can radically transform individuals and organizations*. New York: Currency.

Price, D. J. D., & Gürsey, S. (1975). Studies in scientometrics I: Transience and continuance in scientific authorship. *Ci. Inf. Rio de Janeiro, 4*(1), 27-40.

Rahkovsky, I., Toney, A., Boyack, K. W., Klavans, R., & Murdick, D. A. (2021). AI research funding portfolios and extreme growth. *Frontiers in Research Metrics and Analytics, 6*, 630124. doi:10.3389/frma.2021.630124

Roberts, E. B., & Fusfeld, A. R. (1980). *Critical functions: Needed roles in the innovation process*. Working Paper Alfred P. Sloan School of Management.

Sauermann, H., & Haeussler, C. (2017). Authorship and contribution disclosures. *Science Advances, 3*(11), e1700404.

Spruijt, P., Knol, A. B., Vasileiadou, E., Devilee, J., Lebret, E., & Petersen, A. C. (2014). Roles of scientists as policy advisers on complex issues: A literature review. *Environmental Science & Policy, 40*, 16-25.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports, 9*, 5233. doi:10.1038/s41598-019-41695-z

Wildgaard, L., Schneider, J. W., & Larsen, B. (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics, 101*(1), 125-158.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*, 1036-1039. doi:10.1126/science.1136099

A model for Custom Bibliographic Databases Creation: Machine Learning Approach for Analogue Documents Inclusion

Oskar Kosch[1] and Marek Szarucki[2]

[1] *koscho@uek.krakow.pl*
Cracow University of Economics, Rakowicka 27, 31-510 Krakow (Poland)

[2] *szaruckm@uek.krakow.pl*
Cracow University of Economics, Rakowicka 27, 31-510 Krakow (Poland)

Abstract

To fully take advantage of rigorous literature-based study (e.g., bibliometrics or systematic literature reviews), databases that enable the inclusion of all relevant documents are needed. As many currently available bibliographic databases inherit characteristics of the core-periphery model of scientific production or are unbalanced in their thematic structure, the need for a model for custom bibliographic databases creation procedure arises. At the same time, the adoption of machine learning is apparent throughout science, technology and business, as we learn how to maximise its usefulness. We investigate what model for custom bibliographic database could be adopted, especially taking into consideration the inclusion of analogue publications. Such documents need to be converted into digital form, hence we propose machine learning algorithms and then we explore their accuracy in analogue documents inclusion into the custom bibliographic database for the case of Polish management sciences methodology. We explore the application of neural networks for image preprocessing and optical character recognition, and subsequent application of conditional random fields to obtain a bibliographic database. Our results are highly suggestive and reveal the applicability of the proposed model consisting of database search and snowballing.

Introduction

Web of Science, Scopus, Google Scholar, and beyond

Bibliographic data sources are crucial both for systematic literature reviews (SLRs) and bibliometrics. Their growing importance led to multiple papers on databases selection for both bibliometrics (e.g. entire issue of Quantitative Science Studies, see: Waltman & Larivière, 2020) and SLRs (e.g., Bramer, Rethlefsen, Kleijnen, & Franco, 2017). While bibliometrics usually goes after raw coverage, SLRs do involve other factors, like precision, recall, and number needed to read (NNR); these were the criteria used for database combinations evaluation (e.g., Bramer et al., 2017).

We adapt the definition of a bibliographic database as in the Sile et al. (2018, p. 311): "structured set of bibliographic metadata (e.g. title, publication type, year, and author) in line with requirements for data when calculating the most basic indicator of research output", bearing in mind, that custom databases provide an opportunity to study different types of research output.

Quantitative Science Studies (QSS) dedicated the special issue on data sources (Waltman & Larivière, 2020). It featured Web of Science (WoS, Birkle, Pendlebury, Schnell, & Adams, 2020), Scopus (Baas, Schotten, Plume, Côté, & Karimi, 2020), Dimensions (Herzog, Hook, & Konkiel, 2020), Microsoft Academic Graph (Herzog et al., 2020), CrossRef (Hendricks, Tkaczyk, Lin, & Feeney, 2020), and OpenCitations (Peroni & Shotton, 2020). The issue does not include national databases (which usually contain non-English documents), despite their numerous existence. Sile et al. (2018) listed 21 national European (and Israeli) databases and provided a comprehensiveness study on 13 of them. They have argued, while databases' comprehensiveness may lead to different insights from bibliometrics-supported research evaluation, the main issue remains within the setup of the database (Sile et al., 2018, p. 320). For this reason, conscious data source selection and its further preparation are of such importance. Alternative data sources should be established, as databases not always provide out-of-the-box coverage required; avoiding the inclusion of documents of certain type or language may lead to overlooking current important issues, that of time being are only published in these excluded types or languages. Of course, those important scientific achievements are likely to be published as articles in English, nevertheless, some important details may be missing or waiting till their translation and publication in international journals.

How much may we believe the databases?

The issue of data sources for bibliometrics and systematic literature reviews are far beyond the usual concerns expressed in studies utilising these methods. While usually the limitations of the study that stems from the database imperfections are mentioned, they rarely depict the complexity and choices that are made with the database selection and preparation, raising a question if researchers do not *believe* in the databases integrity inadequately to the reality. Below we present some of the problems, related to data sources in the SLRs and bibliometrics. Franceschini, Maisano, and Mastrogiacomo (2016a) listed errors of the Scopus database, that were later on commented by Elsevier (Meester, Colledge, & Dyas, 2016) to fall into two categories: 1) conservative linking, 2) conservative deduplication. Conservative, that is – precision is more valued than recall. Franceschini, Maisano, and Mastrogiacomo (2016b) further presented their detailed systematic error classification for both Scopus and Web of Science, based on two categories of errors: 1) pre-existing errors in data, and 2) database mapping errors.

Multiple examples from different fields provide a strong basis to argue, that databases are not neutral, and results may vary – for that reason each of the databases might give us a different result (Bar-Ilan, 2017). Therefore, however custom databases might include bias, this bias nevertheless has already been part of existing databases.

Reasons to introduce a model for custom bibliographic database creation procedure

There are examples of using a custom created database to perform context-embedded research: Tunger and Eulerich (2018) used "Jourqual" journal ranking as the basis for the journal selection, from which articles on corporate governance in German-speaking countries were searched. There are several reasons clarifying the need for custom databases, mostly related to the documents' coverage of existing solutions. Web of Science was discussed to be inclined towards English-language journals by 20-25% (Archambault, Vignola-Gagné, Côté, Larivière, & Gingrasb, 2006). De Moya-Anegón et al. (2007) observed that 15% of journals published in Scopus were in a language other than English, compared to 26% of Ulrich's Core collection in 2005. Cowhitt and Cutts (2020), studied thirteen different databases' journals lists and found that databases included in their overlap study "overwhelmingly" cover USA and UK publications' production. These findings are in line with the core-periphery model of scientific production.

Frandsen and Nicolaisen (2008) showed for the case of economics and psychology, there might be not only differences between disciplines but also intradisciplinary ones. This stresses the need for a model for custom database creation procedure not only to include a national perspective but also conduct in-depth disciplines mapping.

Database selection usually restrains the type of documents analysed, and the questions on other types of publications appear. Butler and Visser (2006) explored that a number of "non-source" items, including books and books' chapters, could be found in the WoS (then ISI) database, but they were cautious in their conclusions. Abrizah and Thelwall (2014) researched arts, humanities and social sciences books of five Malaysian university presses and concluded that in total 45% of books were ever cited, with a natural inclination towards older publications. It would mean that 55% of publications cannot be discovered through cited references of the retrieved search results. Hence, the mentioned sources might not be appropriate for their context and language bias, inter- and intra-discipline differences, and selection of the included documents' type. Creating custom databases might overcome some of these issues. It is crucial especially for social sciences and humanities since a significant share of documents is published in national journals, or as other publication types than articles (Sivertsen & Larsen, 2012).

A summary on the development of scientometrics, regarding databases as well, has been recently given by van Raan (2019). When additional sources appeared, the popularity of the research utilising them grew, but also made many studies predicated on the use of such databases. Now, with the ability to automate tasks traditionally done by humans (for the impression of what machines are capable of now, see Wang et al., 2020), the opportunity to create high-quality, custom databases more easily appears. The adoption of machine learning is apparent throughout science, technology and business, as we learn how to maximise its usefulness (Jordan & Mitchell, 2015). We propose a model for custom database creation procedure in social sciences that employ machine learning (ML) algorithms, as this approach outperforms non-ML solutions for citation parsing (Tkaczyk, Collins, Sheridan, & Beel, 2018). Machine learning is a field devoted to "algorithms and techniques for automating solutions to complex problems" (Rebala, Ravi, & Churiwala, 2019), and these algorithms take data to learn a model (rules) of operation.

Providing a model for country-level bibliometrics database creation procedure addresses the call for methodological guidance on the SLRs, and bibliometrics in management (Breslin & Bailey, 2020), but as discussed, the entire social sciences should benefit. We follow the call to reveal methodological issues and research methods relevant to language and country-level context under the cultural influence (Lee, 2020, p. 283), to shed light on "different ways of knowing" (Botha, 2011), to encourage (qualitative) methodology diversity (Bansal & Corley, 2011; Cassell, 2016). Establishing a model that allows for robust country-level analysis could contribute to the plurality of methods available. Creating such a database is not necessarily associated with any nation or country; rather, it should be considered to be custom, tailored to the needs of the research. The research question, that will be further investigated in the next sections, is: **What model for custom bibliographic database creation for non-English documents could be adopted?**

Model for custom bibliographic database creation procedure

As a starting point for custom database creation, we propose a usual search done with existing sources. Echchakoui (2020) argued that database merging is needed, as a combined dataset leads to different results than each of the sources would suggest when taken separately. This is in line with the findings of Bramer et al. (2017). Hence, the use of multiple databases seems reasonable, as the first way of ensuring proper coverage, along with the use of more specialised databases.

The second step, which was proven to be useful in literature search for SLRs, and certainly in the case of bibliometrics as well, is snowballing. Backward snowballing means searching within bibliographic sections of already obtained documents (called "seed documents") and forward snowballing refers to searching for documents citing the ones already obtained (Mourão et al., 2020). We propose a model of procedure that combines the usual search and merging researchers do while conducting systematic literature review or bibliometrics with some postulates of Mourão et al. (2020) mixed search strategy ("DB search + BS*FS"). Our unique contribution is by discussing the need to use less popular (e.g., country-level) databases to provide documents embedded within context (at least, in social sciences) and to showcase pipeline to: 1) scan, 2) post-process, 3) extract and segment citations as needed for further backward snowballing, utilising machine learning. The model for the custom bibliographic database creation procedure is depicted in Figure 1.

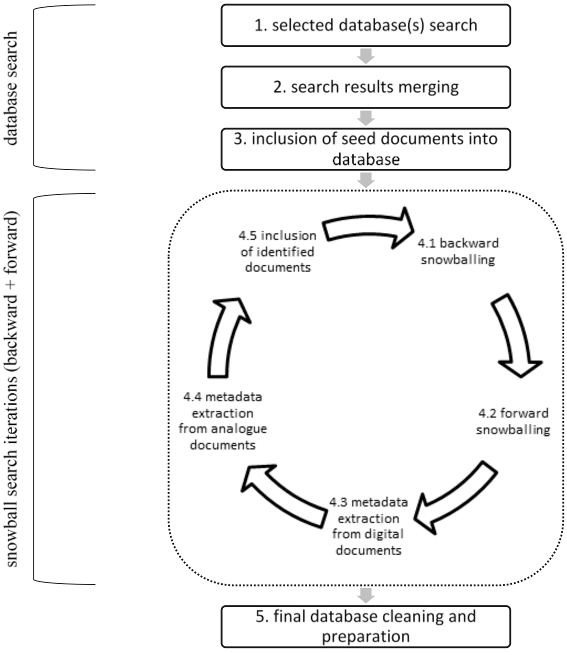

database search
snowball search iterations (backward + forward)
1. selected database(s) search
2. search results merging
3. inclusion of seed documents into database
4.1 backward snowballing
4.2 forward snowballing
4.3 metadata extraction from digital documents
4.4 metadata extraction from analogue documents
4.5 inclusion of identified documents
5. final database cleaning and preparation

Figure 1. Model of the custom bibliographic database creation procedure
Source: Own study based on Echchakoui (2020), Bramer et al. (2017), and Mourão et al. (2020).

We aim to explore the applicability of the above model to create a custom bibliographic database for social sciences research of non-English documents. The model itself should be applicable not only within the case studied. For the reasons outlined in the introduction, due to context-dependency, we believe the most significant impact it may achieve in social sciences. In this paper, we provide qualitative research of exploratory type for the case of the Polish management methodology scholars, as this is a perfect example of a research object for which bibliometric/systematic literature review would be extremely difficult without the use of the custom created bibliographic database. The details of the methodology used are presented in the subsequent section.

Methodology

This study is qualitative research of exploratory type – it tries to answer the question if a database could be created for the case given, with the theoretical background presented in the prior section. We argue that further efforts will allow for establishing a robust model to both preserve rigour of SLRs and bibliometrics while bringing diversity that custom bibliographic databases might offer.

For the case studied in this paper, we applied only backward snowballing, for the literature studied in majority comprises publications absent from citation indexing databases. Also, the forward snowballing is relatively well-described, e.g., by Felizardo, Mendes, Kalinowski, Souza, and Vijaykumar (2016).

Despite the theoretical contribution, our unique approach is exemplified by the use of machine learning for step 4.4 of the proposed model. While the execution of other steps is well-discussed in articles, the extraction of relevant metadata from printed documents remains to be challenging and unapproached. Several recommendations for computer vision appliances for books processing and scanning (Prashanth, Sai Vivek, Ruthvik Reddy, & Aruna, 2019;

Suddaby, Ganzin, & Minkus, 2017), even these in braille (Kawabe, Seto, Nambo, & Shimomura, 2020), are available. The first section ("3.1. Database search") will present the databases search we utilised, while the second will discuss the mentioned snowball search with the use of the machine learning approach ("3.2. Snowball search").

Database search

Three databases were queried several times, with the last run being made on 10.12.2020: Web of Science, Scopus, BazEkon[1].

The queries were following:

1. Scopus: authkey("research method*" or methodology) or title("research method*" or methodology) and subjarea(busi) and affilcountry(poland) and doctype(ar)
2. Web of Science: ((ak=("research method*" or methodology) or ti=("research method*" or methodology)) and su=(business & economics) and cu=(poland)) and document types: (article)
3. BazEkon: sl_kl = metodologia badań or sl_kl = metody badawcze or sl_kl = methodology of science or sl_kl = research methodology or sl_kl = research methods

BazEkon is focused on the fields of science equivalent to the Scopus "BUSI" and WoS "Business & Economics". The further abstract screening was performed as part of the inclusion strategy.

Snowball search

The necessity to perform a large number of scans (not all documents feature a bibliography section, as they provide footnotes; also, our project includes future content analysis), the use of standard scanners would be too time-consuming. Cutting publications to use fast scanning was not possible either. Hence, we decided to take an approach known from different hardware settings (e.g., BookLiberator, n.d.) and software approaches (e.g., Quan, Zhou, & Chen, 2017) to take photos and process them further. Image processing to obtain extracted pages and perform optical character recognition is a well-described task and there is little innovation to bring. Nevertheless, when the number of pages needed to scan for analysis counts in tens of thousands (this project involves ca. 80 000 scanned pages) the process itself (pipeline) has to become robust and for this simple reason deserves description. Another restriction was cost-effectiveness; the research project itself did not involve financing high enough to obtain any of advanced scanning machines. Instead, simple hardware tool was developed, with total costs below $35. The hardware requires a camera or smartphone to take photos of both pages at once. We used a relatively small dataset of 200 segmented photos of books for Tensorflow (Abadi et al., 2015) U-Net convolutional neural network (Ronneberger, Fischer, & Brox, 2015) to extract pages. Further, we have performed perspective transformation and cleaning using OpenCV (Bradski, 2000) python library. Next in our pipeline was Tesseract OCR (tesseract-ocr, 2020) to perform optical character recognition (OCR). For the reasons of modularity, future extension, and updates, we decided to use AnyStyle (inukshuk, 2020) as our last step, that is, metadata cited references parsing. One of the full-featured alternatives to use with born-digital documents is GROBID (kermitt2, 2020), which is also actively maintained. Unfortunately, CERMINE (CeON, 2018; Tkaczyk, Szostek, Fedoryszak, Dendek, & Bolikowski, 2015) the last update on the GitHub repo was done in 2018, which puts in question further use of this tool. Both of them did not perform well on scanned documents. We decided to use all the BazEkon seed documents to retrain the AnyStyle parser; the default model displayed fair performance with English-language, but for the context studied needed training on custom data. We evaluate the impact

[1] https://bazekon.uek.krakow.pl

of training data volume and variety by random splitting dataset and hence performing accuracy simulation (50 samples for each level of volume, and variety were collected, with 1530 training and 382 test references manually segmented). Employing this approach should inform us on the desired size of the training dataset for the tool used.

Machine learning is not the only option, but the most flexible one to perform pages extraction and OCR; various digitalisation programs and documentation management tools, like ABBYY FineReader or Readiris, allow for photos processing, but few customizations can be done, comparing to those available when programming in R or python.

For data merging, we used a citation key for exact matching, similar to the one provided by the Web of Science plain text export.

Results

Databases search

After inclusion was performed, basic descriptive statistics of databases involved were calculated (see Table 1). Scopus usually is a better solution for social sciences, and it brought more results, especially for articles involving European affiliation than Web of Science. However, WoS proved to be better for the discovery of the Polish methodology articles – first, it provided a greater share of Poland-affiliated to European-affiliated documents than Scopus did, finally bringing almost twice more records; secondly, the number needed to read was significantly lower, and the number of results included almost three times more than in Scopus. After deduplication, it was discovered, that all articles included in the Scopus database, were also found in Web of Science.

Table 1. Database search results and inclusion.

Variable	*Scopus*	*WoS*	*BazEkon*
Worldwide articles (W)	8017	6755	-
Europe-affiliated articles (E)	3694	2432	-
Poland-affiliated (P)	68	112	905
E/W	46.08%	36.00%	-
P/W	0.85%	1.66%	-
P/E	1.84%	4.61%	-
Included articles	6	22	81
Number Needed to Read (NNR)	11.33	5.09	11.17

The only leverage in terms of non-source items, that could be calculated precisely, was the one concerning Journal Citation Report (JCR) items abbreviations (Clarivate, n.d.) as the abbreviations list (Web of Knowledge, n.d.) has shown deficiencies in retrieving abbreviations of items, including those in Emerging Sources Citation. For a merged WoS-Scopus database, we extracted 1081 references (765 unique), of which 409 (194 unique) pointed JCR sources, and 672 (571 unique) pointed non-JCR sources. Therefore, for unique items, the non-JCR share was 74.64%.

Snowball search – analogue document metadata extraction

AnyStyle, utilising Conditional Random Fields (CRF) machine learning approach provides researchers with state-of-the-art citations segmentation output. One drawback of the machine learning approach employed with AnyStyle needs to supply manually segmented references which will serve as the training dataset. It is not clear how many manually segmented samples should be provided (the volume of a dataset), and from how many distinct documents (variety of dataset). Hence, some basic analysis of the accuracy improvement while increasing variety will be performed using both simulation and slicing dataset. For one of the random splits, we present overall performance in Table 2. As tested, the valid keys counted for 358 out of 382 test dataset, resulting in 93.72% accuracy of the merged author, year and title field.

Table 2. Cited references parsing performance.

field	*precision*	*recall*	*F1*	*accuracy*
author	0.991	0.995	0.993	0.986
date	0.982	0.979	0.981	0.995
title	0.991	0.987	0.989	0.958
publisher	0.973	0.979	0.976	0.961
location	0.974	0.971	0.972	0.970
container title	0.979	0.970	0.974	0.902
editor	0.950	0.975	0.962	0.897
journal	0.982	0.975	0.978	0.978
volume	0.983	0.988	0.986	0.937
citation number	1.000	0.997	0.999	0.992
pages	1.000	1.000	1.000	1.000
URL	1.000	1.000	1.000	1.000
edition	0.995	0.987	0.991	1.000
note	0.990	0.972	0.981	0.500
DOI	0.997	0.995	0.996	1.000

It is seen (Figure 2) that a dataset exceeding 250 sequences is mostly important to improve the whole sequence segmentation although, n terms of individual tokens (words) the change is less impressive. Thus, for single field analysis (e.g., this of cited authors) the segmented dataset might start as low as 250 segmentations, but if more fields are used at once (e.g., as composite IDs for document citation-based network creation) the required dataset size should not be smaller than 1500 segmented references (to which testing dataset should be added, with at least 20% of training volume). To the issue of volume and variety of dataset, the problem of the structure of citations is added.

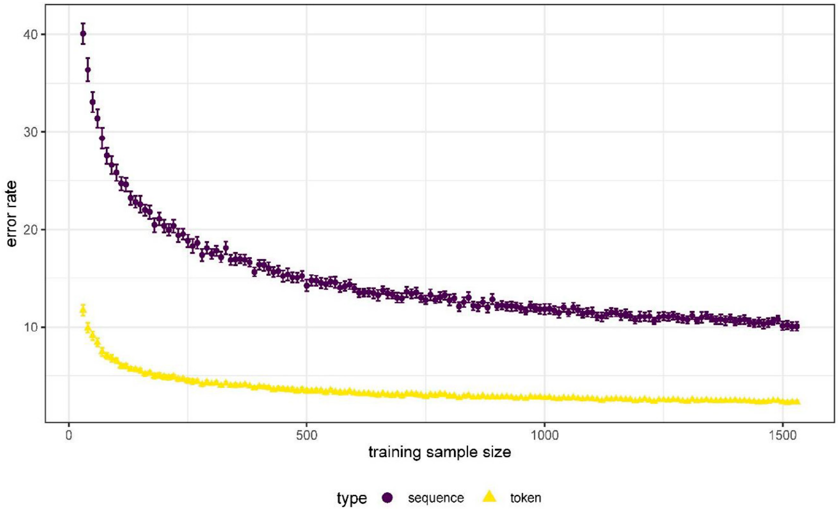


Figure 2. Training sample size (volume) vs error rate.
Source: own study.

Dataset variety significantly influences the error rate (see Figure 3). Researchers should aim for increasing variety of dataset, and only then increase its volume.

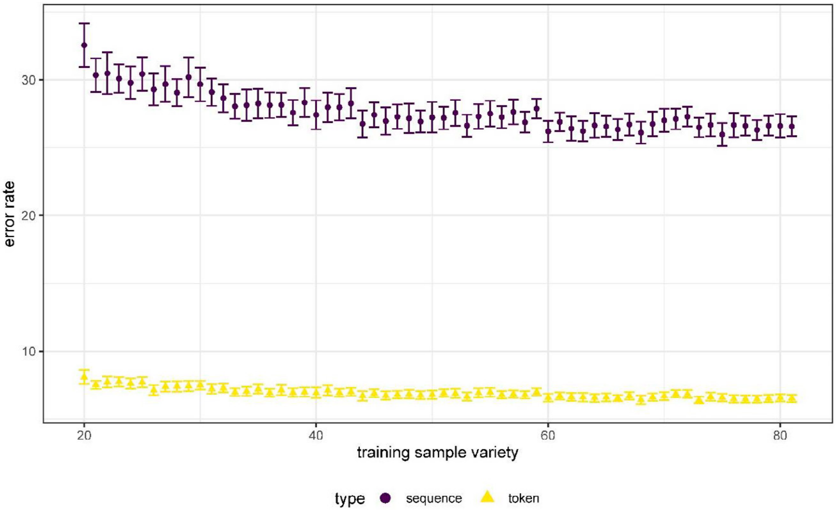
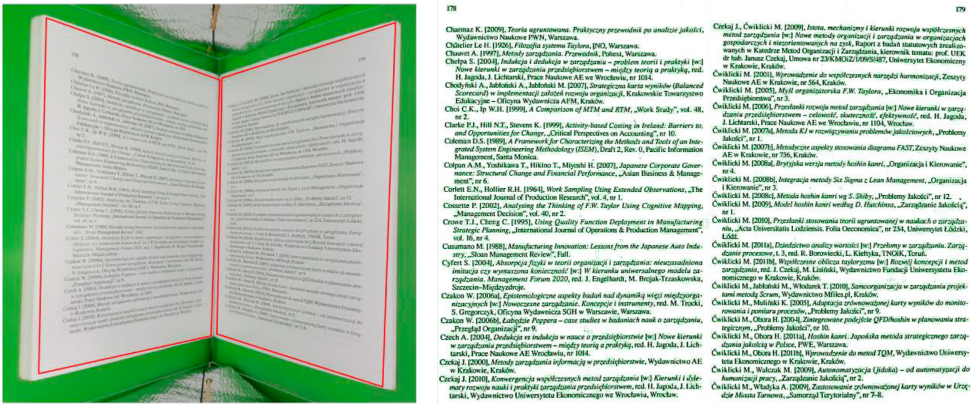


Figure 3. Training sample variety vs error rate.
Source: own study.

Scanning itself takes about 1-4 seconds per page, depending on the condition of the book and the type of paper being used. Cited references creation is therefore quite fast, for bibliography usually spans across few pages only. Tesseract OCR performed fairly (however, was resource-consuming in terms of time and computing, for results see Figure 4), but the text overlay (at the current state of the project) did not allow for automatic references extraction with satisfying accuracy. Instead, manual treatment (taking on average about three minutes per document) was needed. Subsequent cited references parsing performed well.



178

Charmaz K. [2009], *Teoria ugruntowana. Praktyczny przewodnik po analizie jakości*, Wydawnictwo Naukowe PWN, Warszawa.
Châtelier Le H. [1926], *Filozofia systemu Taylora*, [NO, Warszawa.
Chauvet A. [1997], *Metody zarządzania. Przewodnik*, Poltext, Warszawa.
Chełpa S. [2004], *Indukcja i dedukcja w zarządzaniu – problem teorii i praktyki* [w:] *Nowe kierunki w zarządzaniu przedsiębiorstwem – między teorią a praktyką*, red. H. Jagoda, J. Lichtarski, Prace Naukowe AE we Wrocławiu, nr 1014.
Chodyński A., Jabłoński A., Jabłoński M. [2007], *Strategiczna karta wyników (Balanced Scorecard) w implementacji założeń rozwoju organizacji*, Krakowskie Towarzystwo Edukacyjne – Oficyna Wydawnicza AFM, Kraków.
Choi C.K., Ip W.H. [1999], *A Comparison of MTM and RTM*, „Work Study", vol. 48, nr 2.
Clarke P.J., Hill N.T., Stevens K. [1999], *Activity-based Costing in Ireland: Barriers to, and Opportunities for Change*, „Critical Perspectives on Accounting", nr 10.
Coleman D.S. [1989], *A Framework for Characterizing the Methods and Tools of an Integrated System Engineering Methodology (ISEM)*, Draft 2, Rev. 0, Pacific Information Management, Santa Monica.
Colpan A.M., Yoshikawa T., Hikino T., Miyoshi H. [2007], *Japanese Corporate Governance: Structural Change and Financial Performance*, „Asian Business & Management", nr 6.
Corlett E.N., Hollier R.H. [1964], *Work Sampling Using Extended Observations*, „The International Journal of Production Research", vol. 4, nr 1.
Cossette P. [2002], *Analysing the Thinking of F.W. Taylor Using Cognitive Mapping*, „Management Decision", vol. 40, nr 2.
Crowe T.J., Cheng C. [1995], *Using Quality Function Deployment in Manufacturing Strategic Planning*, „International Journal of Operations & Production Management", vol. 16, nr 4.
Cusumano M. [1988], *Manufacturing Innovation: Lessons from the Japanese Auto Industry*, „Sloan Management Review", Fall.
Cyfert S. [2004], *Absorpcja fizyki w teorii organizacji i zarządzania: nieuzasadniona imitacja czy wymuszona konieczność* [w:] *W kierunku uniwersalnego modelu zarządzania. Management Forum 2020*, red. J. Engelhardt, M. Brojak-Trzaskowska, Szczecin–Międzyzdroje.
Czakon W. [2006a], *Epistemologiczne aspekty badań nad dynamiką więzi międzyorganizacyjnych* [w:] *Nowoczesne zarządzanie. Koncepcje i instrumenty*, red. M. Trocki, S. Gregorczyk, Oficyna Wydawnicza SGH w Warszawie, Warszawa.
Czakon W. [2006b], *Łabędzie Poppera – case studies w badaniach nauk o zarządzaniu*, „Przegląd Organizacji", nr 9.
Czech A. [2004], *Dedukcja vs indukcja w nauce o przedsiębiorstwie* [w:] *Nowe kierunki w zarządzaniu przedsiębiorstwem – między teorią a praktyką*, red. H. Jagoda, J. Lichtarski, Prace Naukowe AE Wrocławiu, nr 1014.
Czekaj J. [2000], *Metody zarządzania informacją w przedsiębiorstwie*, Wydawnictwo AE w Krakowie, Kraków.
Czekaj J. [2010], *Konwergencja współczesnych metod zarządzania* [w:] *Kierunki i dylematy rozwoju nauki i praktyki zarządzania przedsiębiorstwem*, red. H. Jagoda, J. Lichtarski, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

179

Czekaj J., Ćwiklicki M. [2009], *Istota, mechanizmy i kierunki rozwoju współczesnych metod zarządzania* [w:] *Nowe metody organizacji i zarządzania w organizacjach gospodarczych i niezorientowanych na zysk*, Raport z badań statutowych zrealizowanych w Katedrze Metod Organizacji i Zarządzania, kierownik tematu: prof. UEK dr hab. Janusz Czekaj, Umowa nr 23/KMOiZ/1/09/S/487, Uniwersytet Ekonomiczny w Krakowie, Kraków.
Ćwiklicki M. [2001], *Wprowadzenie do współczesnych narzędzi harmonizacji*, Zeszyty Naukowe AE w Krakowie, nr 564, Kraków.
Ćwiklicki M. [2005], *Myśl organizatorska F.W. Taylora*, „Ekonomika i Organizacja Przedsiębiorstwa", nr 3.
Ćwiklicki M. [2006], *Przesłanki rozwoju metod zarządzania* [w:] *Nowe kierunki w zarządzaniu przedsiębiorstwem – celowość, skuteczność, efektywność*, red. H. Jagoda, J. Lichtarski, Prace Naukowe AE we Wrocławiu, nr 1104, Wrocław.
Ćwiklicki M. [2007a], *Metoda KJ w rozwiązywaniu problemów jakościowych*, „Problemy Jakości", nr 1.
Ćwiklicki M. [2007b], *Metodyczne aspekty stosowania diagramu FAST*, Zeszyty Naukowe AE w Krakowie, nr 736, Kraków.
Ćwiklicki M. [2008a], *Brytyjska wersja metody hoshin kanri*, „Organizacja i Kierowanie", nr 4.
Ćwiklicki M. [2008b], *Integracja metody Six Sigma z Lean Management*, „Organizacja i Kierowanie", nr 3.
Ćwiklicki M. [2008c], *Metoda hoshin kanri wg S. Shiby*, „Problemy Jakości", nr 12.
Ćwiklicki M. [2009], *Model hoshin kanri według D. Hutchinsa*, „Zarządzanie Jakością", nr 1.
Ćwiklicki M. [2010], *Przesłanki stosowania teorii ugruntowanej w naukach o zarządzaniu*, „Acta Universitatis Lodziensis. Folia Oeconomica", nr 234, Uniwersytet Łódzki, Łódź.
Ćwiklicki M. [2011a], *Dziedzictwo analizy wartości* [w:] *Przełomy w zarządzaniu. Zarządzanie procesowe*, t. 3, red. R. Borowiecki, L. Kiełtyka, TNOiK, Toruń.
Ćwiklicki M. [2011b], *Współczesne oblicza tayloryzmu* [w:] *Rozwój koncepcji i metod zarządzania*, red. J. Czekaj, M. Lisiński, Wydawnictwo Fundacji Uniwersytetu Ekonomicznego w Krakowie, Kraków.
Ćwiklicki M., Jabłoński M., Włodarek T. [2010], *Samoorganizacja w zarządzaniu projektami metodą Scrum*, Wydawnictwo Mfiles.pl, Kraków.
Ćwiklicki M., Muliński K. [2005], *Adaptacja zrównoważonej karty wyników do monitorowania i pomiaru procesów*, „Problemy Jakości", nr 9.
Ćwiklicki M., Obora H. [2004], *Zintegrowane podejście QFD/hoshin w planowaniu strategicznym*, „Problemy Jakości", nr 10.
Ćwiklicki M., Obora H. [2011a], *Hoshin kanri. Japońska metoda strategicznego zarządzania jakością w Polsce*, PWE, Warszawa.
Ćwiklicki M., Obora H. [2011b], *Wprowadzenie do metod TQM*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.
Ćwiklicki M., Walczak M. [2009], *Autonomatyzacja (jidoka) – od automatyzacji do humanizacji pracy*, „Zarządzanie Jakością", nr 2.
Ćwiklicki M., Władyka A. [2009], *Zastosowanie zrównoważonej karty wyników w Urzędzie Miasta Tarnowa*, „Samorząd Terytorialny", nr 7–8.

Figure 4. Page trapezoid detection (left, trapezoids on pages) and Optical Character Recognition bounding boxes of possible characters on cleaned pages (right, boxes).
Source: own study.

Conclusions

The main conclusion, as depicted by performing the most controversial step of custom bibliographic database creation, “4.4 metadata extraction from analogue documents”, is that the proposed model was successfully applied, and should be further developed. Hence, the model for custom bibliographic database creation utilising database search for seed documents and subsequent backward and forward snowballing, including analogue documents as well, answers our research question. Both pages extraction and cited references parsing presented fair performance, hence allowing to include analogue documents and gain independence from the selected coverage of existing bibliographic databases. The proposed model of custom bibliographic database creation proved to be useful in the case studied. Hopefully, it could contribute to the more diversified systematic literature reviews and bibliometrics in all fields of social sciences, when conducted for non-English context – however, we only attempted to employ the model for a single case, as this was exploratory research, but we encourage other researchers to test our model in different settings or contribute with their own.

The last conclusion is related to the cited references’ parser training dataset creation. 1500 references from at least 100 unique sources should give expected results for a more complex study, while for single field extraction even 250 might work. Better than segmenting a few articles’ full references, it is to sample a couple of references out of each document, to achieve higher variety and overall accuracy.

Limitations and recommendations for future research

We deliberately have not included a part for forward snowballing, as it is described by other researchers, and the procedure similar to one presented by Felizardo, Mendes, Kalinowski, Souza, and Vijaykumar (2016) could be used. Optimally, after each step of backward snowballing, forward snowballing should be carried out – but in our case, the majority of included literature is absent from citation indexing databases, hence the utility of providing forward snowballing was low.

Certain topics require strict inclusion criteria for the reason of the relative popularity of keywords as compared against the actual content of the paper. Possibly full-text analysis utilising e.g., machine learning classification trained on agreed dataset could serve as more objectified filtering than doing it manually. It should form a basis for future research recommendations.

Other limitations stem from the training datasets; enlarging cited references data volume and variety should increase the accuracy of the model, and the same applies to the U-Net convolutional neural network. While the current performance was fair, using better models allows for more precise research output, and before undertaking the next steps in the conducted research project, we should try to train such models.

Also, models for books’ sections extraction and citations attribution should be constructed, to allow for a study of research items of less variance in length (in terms of pages). Precise references extraction models should be proposed as part of the pipeline mentioned, to further reduce human input in the database creation. Preferably these should result in software that takes images as input and provides bibliographic records as output.

Some limitations of the proposed model itself stem from technological barriers and data accuracy. However using machine learning contributes to a significant reduction of time and effort needed for custom database creation, the process of scanning analogue documents is still time-consuming and can damage original documents. Also, quality control of the data is needed, as in the case of extant databases; extensive post-processing eliminating known issues (e.g., names disambiguation) might introduce serious challenge for researchers, especially given different alphabets and naming conventions.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. Retrieved January 10, 2021, from http://arxiv.org/abs/1603.04467

Abrizah, A., & Thelwall, M. (2014). Can the impact of non-Western academic books be measured? An investigation of Google Books and Google Scholar for Malaysia. *Journal of the Association for Information Science and Technology*, 65(12), 2498–2508.

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingrasb, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, 68(3), 329–342.

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, 1(1), 377–386.

Bansal, P., & Corley, K. (2011, April 1). From the editors the coming of age for qualitative research: Embracing the diversity of qualitative methods. *Academy of Management Journal*.

Bar-Ilan, J. (2017). *Bibliometrics of "information retrieval" - A tale of three databases*. *CEUR Workshop Proceedings* (Vol. 1888). CEUR-WS.

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, 1(1), 363–376.

BookLiberator. (n.d.). *BookLiberator*. Retrieved January 12, 2021, from http://bookliberator.com/

Botha, L. (2011). Mixing methods as a process towards indigenous methodologies. *International Journal of Social Research Methodology*, 14(4), 313–325.

Bradski, G. (2000). The OpenCV Library. *Dr Dobbs Journal of Software Tools*, 25, 120–125.

Bramer, W. M., Rethlefsen, M. L., Kleijnen, J., & Franco, O. H. (2017). Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study. *Systematic Reviews*, 6(245).

Breslin, D., & Bailey, K. (2020). Expanding the Conversation through 'Debate Essays' and 'Review Methodology' Papers. *International Journal of Management Reviews*, 22(3), 219–221.

Butler, L., & Visser, M. S. (2006). Extending citation analysis to non-source items. *Scientometrics*, 66(2), 327–343.

Cassell, C. (2016). European qualitative research: A celebration of diversity and a cautionary tale. *European Management Journal*, 34(5), 453–456.

CeON. (2018). *CeON/CERMINE: Content ExtRactor and MINEr*. Retrieved January 11, 2021, from https://github.com/CeON/CERMINE

Clarivate. (n.d.). *Web of Science Master Journal List - Collection List Downloads*. Retrieved January 11, 2021, from https://mjl.clarivate.com/collection-list-downloads

Cowhitt, T., & Cutts, A. (2020). Using network analysis to compare bibliographic database journal coverage. *Journal of Electronic Resources Librarianship*, 32(3), 195–210.

Echchakoui, S. (2020). Why and how to merge Scopus and Web of Science during bibliometric analysis: the case of sales force literature from 1912 to 2019. *Journal of Marketing Analytics*, 8(3), 165–184.

Felizardo, K. R., Mendes, E., Kalinowski, M., Souza, É. F., & Vijaykumar, N. L. (2016). Using Forward Snowballing to update Systematic Reviews in Software Engineering. *International Symposium on Empirical Software Engineering and Measurement* (Vol. 08-09-September-2016, pp. 1–6).

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016a). The museum of errors/horrors in Scopus. *Journal of Informetrics*, 10(1), 174–182.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016b). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics*, 10(4), 933–953.

Frandsen, T. F., & Nicolaisen, J. (2008). Intradisciplinary differences in database coverage and the consequences for bibliometric research. *Journal of the American Society for Information Science and Technology*, 59(10), 1570–1581.

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1), 414–427.

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, 1(1), 387–395.

inukshuk. (2020). *inukshuk/anystyle: Fast and smart citation reference parsing*. Retrieved January 12, 2021, from https://github.com/inukshuk/anystyle

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260.

Kawabe, H., Seto, S., Nambo, H., & Shimomura, Y. (2020). Experimental study on scanning of degraded braille books for recognition of dots by machine learning. *Advances in Intelligent Systems and Computing* (Vol. 1001, pp. 322–334).

kermitt2. (2020). *kermitt2/grobid: A machine learning software for extracting information from scholarly documents*. Retrieved January 11, 2021, from https://github.com/kermitt2/grobid

Lee, B. (2020, March 1). Methodology Matters; Even More. *European Management Review*.

Meester, W. J. N., Colledge, L., & Dyas, E. E. (2016). A response to "The museum of errors/horrors in Scopus" by Franceschini et al. *Journal of Informetrics*, 10(2), 569–570.

Mourão, E., Pimentel, J. F., Murta, L., Kalinowski, M., Mendes, E., & Wohlin, C. (2020). On the performance of hybrid search strategies for systematic literature reviews in software engineering. *Information and Software Technology*, 123, 106294.

De Moya-Anegón, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., Muñoz-Fernández, F. J., González-Molina, A., & Herrero-Solana, V. (2007). Coverage analysis of Scopus: A journal metric approach. *Scientometrics*, 73(1), 53–78.

Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, 1(1), 428–444.

Prashanth, P., Sai Vivek, K., Ruthvik Reddy, D., & Aruna, K. (2019). Book detection using deep learning. *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019* (pp. 1167–1169).

Quan, N., Zhou, X., & Chen, S. (2017). Scan paperback books by a camera. *2016 IEEE International Conference on Information and Automation, IEEE ICIA 2016* (pp. 579–584).

van Raan, A. (2019). Measuring science: Basic principles and application of advanced bibliometrics. *Springer Handbook of Science and Technology Indicators* (pp. 237–280). Cham: Springer.

Rebala, G., Ravi, A., & Churiwala, S. (2019). Machine Learning Definition and Basics. *An Introduction to Machine Learning*, 1–17.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9351, pp. 234–241).

Sile, L., Pölönen, J., Sivertsen, G., Guns, R., Engels, T. C. E., Arefiev, P., Dušková, M., et al. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. *Research Evaluation*, 27(4), 310–322.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics*, 91(2), 567–575.

Suddaby, R., Ganzin, M., & Minkus, A. (2017). Craft, magic and the re-enchantment of the world. In S. Siebert (Ed.), *Management Research: European Perspectives* (pp. 41–72). New York: Taylor and Francis.

tesseract-ocr. (2020). *tesseract-ocr/tesseract*. Retrieved January 12, 2021, from https://github.com/tesseract-ocr/tesseract

Tkaczyk, D., Collins, A., Sheridan, P., & Beel, J. (2018). Machine Learning vs. Rules and Out-of-the-Box vs. Retrained: An Evaluation of Open-Source Bibliographic Reference and Citation Parsers. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (Vol. 10, pp. 99–108).

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015). CERMINE: Automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition*, 18(4), 317–335.

Tunger, D., & Eulerich, M. (2018). Bibliometric analysis of corporate governance research in German-speaking countries: applying bibliometrics to business research using a custom-made database. *Scientometrics*, 117(3), 2041–2059.

Waltman, L., & Larivière, V. (2020). Special issue on bibliographic data sources. *Quantitative Science Studies*, 1(1), 360–362.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, 1(1), 396–413.

Web of Knowledge. (n.d.). *Journal Title Abbreviations*. Retrieved January 11, 2021, from https://images.webofknowledge.com/WOKRS535R102/help/WOS/A_abrvjt.html

Publish more or publish differently? New aspects of relationship between scientific mobility and performance of young researchers

Maxim Kotsemir[1], Ekaterina Dyachenko[2] and Alena Nefedova[3]

[1] *mkotsemir@hse.ru*
National Research University Higher School of Economics, Institute for Statistical Studies and Economics of Knowledge, Myasnitskaya 11, 443, Moscow (Russia). Corresponding author.

[2] *dyachenko-el@ranepa.ru*
The Russian Presidential Academy of National Economy and Public Administration (RANEPA), Prospect Vernadskogo 84, bldg. 9, 1204, Moscow (Russia)

[3]*anefedova@hse.ru*
National Research University Higher School of Economics, Institute for Statistical Studies and Economics of Knowledge, Myasnitskaya 11, 443, Moscow (Russia)

Abstract

The paper presents the preliminary results of a research project on the international mobility of young Russian researchers. This study is focused on the impact of education or work experience abroad on their future scientific careers, particularly their publication activity. The case study of the one large Russian university was examined: a unique for Russia database combining both biographical data (information from CV published openly) and publication activity indicators (data from Scopus) of employees of this university was collected. A positive relationship between the international mobility and scientific productivity of scientists was revealed. Those involved in international mobility not only publish more scientific articles on average but their papers are published in journals of higher level and are cited more often. We also found some differences in the choice of topics and collaboration behavior between mobile and non-mobile scientists.

Introduction

Academic mobility is one of the key issues in the S&T policy for the last four decades in developed countries. For nations, mobile researchers are seen as a good source for enhancing their national innovation system. For research organizations, scientists with experience working abroad can be a valuable source for international collaboration. According to the results of previous studies, the experience of studying or working abroad increases involvement in the process of exchange of knowledge and technologies both with other countries and within their own country (De Filippo et al., 2009; Edler et al., 2011; Scellato et al., 2012, 2017). Moreover, the experience of mobility can significantly impact the cognitive career of researcher, bringing new perspectives or new theoretical or empirical approaches; as well as community career, by gaining new links with peers in the scientific community (Gläser & Laudel, 2015), or even impact on the change of research practice (Gläser et al., 2014). However, the effects of mobility can vary on different career stages of researchers and mobility duration (Cañibano et al., 2020) or even scientific discipline (Laudel & Bielick, 2019).

There is a well-known assumption that mobile researchers show higher publication activity. However, the empirical evidence is contradictory so far. Most studies have shown that scientists with international experience have higher research performance (Gureyev et al., 2020; Netz, Hampel & Aman, 2020). Still, there is evidence that mobile researchers failed to demonstrate higher research productivity than their local counterparts (Horta et al., 2019; Shin et al. 2014). There is a possible explanation that this difference comes from national research systems where mobile researchers come back. Within the last 15 years, many studies devoted to returnee's career tracks and their publication activity (those specialists who left the country and then returned). A study based on the example of scientists from Argentina showed that international work experience does not affect the total number of publications, but it does affect the share of publications in high-impact journals (Jonkers & Cruz-Castro, 2013). Similar results were in a study of Chinese researchers (Jonkers & Tijssen, 2008). In addition, an analysis of the career growth of 370 professors in Japan showed that work experience abroad was statistically associated with faster subsequent career advancement, even though there was no effect on publication activity (Lawson & Shibayama, 2015). A study devoted to Chinese mobile researchers showed, returnees are encountering significant barriers in publishing their work in international journals after their change in affiliation from an overseas to a Chinese university (Gao & Liu, 2020). Moreover, they are struggling to change their research focus when they returned to China because the previous one fails to fit the domestic research agenda (Ibid.).

To the best of our knowledge, Russian scientists are not studied thoroughly in this way. The 'brain drain' was the dominant framework during the 1990s, when scientists were leaving the country en masse. For a long time, the mobility of scientists was analyzed only from that perspective. However, in the last decade, the agenda has begun to change. In this study, the case of one large Russian university was examined. The study is focused on the impact of education or work experience abroad on the scientific career of young researchers, particularly on their publication activity.

Following the goal of our study, we compiled the unique for Russia database with the combination of biographical data got from their CVs and bibliometric information derived from the Scopus database. The goal was to overcome the significant incompleteness of the data extracted from bibliography descriptions of publications alone. This approach is not the most common in studies of mobility and research performance, since the data sets required are not easily available. Implementation of such an approach in Russia is even more limited due to specificities of public availability of bibliographical data of Russian scientists. Most Russian universities and research organizations post very brief information about their employees. It is also unusual practice for Russian scientists to publish their CVs in open access.

One of our aims was to look beyond the performance indicators to understand what exactly changes in the research profile of a scientist after he or she moves to another country and then returns. The novelty of our study is as follows: when exploring the relationship between mobility and productivity of researchers we try to reveal the mechanisms of this relation. When it is discovered that mobile researchers publish more or get more citations than non-mobile, sometimes the authors make suggestions that mobility could affect this in several ways – mobile researchers could be more motivated to publish, they probably have a bigger circle of potential coauthors, they gain access to new knowledge and develop new skills in new places, they transfer knowledge themselves, they also can get access to resources and infrastructure. While the plausibility of these factors is confirmed in qualitative studies there is a lack of quantitative research on these factors (Netz, Hampel & Aman, 2020). Today we still do not know what factors are the main drivers of productivity and impact.

In this exploratory study, we analyze several aspects of relationship between mobility and the publication profiles of researchers. The first one is the overall researcher performance. The first hypothesis (H1) is that mobile researchers will have higher number of publications better citation indicators and publish in journals of higher quality. The second one is a collaboration. The second hypothesis (H2) is that mobile researchers tend to have more professional ties which affect productivity. It was tested in some studies with the data on co-authorship but usually was discussed in the context of international collaboration (for example Jonkers & Cruz-Castro, 2013; Jöns, 2007; Netz, Hampel & Aman, 2020). We were interested in whether mobile researchers collaborate more than non-mobile, considering not only international collaboration. The third factor we study here is research topics. The third hypothesis (H3) is that mobile scientists have more diverse research portfolios in terms of research topics. Mobility can cause a shift in the topics researchers are working on, which can affect performance. A researcher can start a new topic in a new place, probably more "fruitful", can drop old ones. To our knowledge, this subject has not been investigated in quantitative studies scientific mobility. In general, mobile researchers not only publish more – they make research and publish differently, and bibliometric analysis can be used to trace some of this difference.

Methodology

For our study, we chose the case of the National Research University Higher School of Economics (HSE University), a large university situated in Moscow with branches in three other cities (Saint-Petersburg, Nizhniy Novgorod, and Perm). Despite the title, HSE University has a very diverse range of departments and research units, from physical to philosophical[1]. HSE University has its own internal programs to support international mobility, is actively working to integrate researchers and teachers into the international academic community, and is recruiting specialists from leading foreign universities and research organizations.

HSE University is among Russian leaders in terms of completeness and availability of data on employees located in the public domain[2]. Each HSE University employee has a personal page on university web-portal with data on education, professional experience, publications etc. including the info on his/her professional/academic/studying mobility. Many HSE employees post their CVs on personal pages. We've collected a database of young mobile and non-mobile researchers affiliated with HSE University where bibliometric data (information about Scopus-indexed publications) were combined with the biographical information of researchers (experience of international mobility) gathered from their personal pages at HSE University web-portal. Among all HSE employees, we've selected those who held a research position as main position marked on the employee's personal page. Further among all researchers we've selected HSE those who as of March 2020, were 39 years old or younger, and who explicitly marked on his/her personal page (or in CV) at least one episode of scientific or educational mobility lasting more than three months. Processing of personal pages and CVs of HSE employees was done in an automatic mode with further manual check and correction of controversial cases. As a result, we have selected the initial sample of 193 young mobile researchers working at HSE University – those who were under 40 year (as of March 2020), held a research position and had at least one episode of international mobility lasting 3 months or more.

To form the group of non-mobile researchers, we've used matched pairs method[3]. The aim was to create two groups for comparison, where members are as close as possible in several significant parameters. For each mobile researcher, a non-mobile "pair" was selected manually from the pool of HSE researchers without mobility experience – a researcher working in the same or close by profile department, of about the same age (the difference in the year of receiving the first diploma was no more than two years). In addition, control was carried out for the region of the first higher education – for mobile researchers who received it in Moscow or St. Petersburg, non-mobile pair with the same educational background were selected. The selection resulted in 119 pairs of mobile and non-mobile researchers.

To analyze whether the length of the mobility episodes makes a difference to research performance, we divided all mobile researchers to two sub-samples – researchers with episodes

[1] See more about HSE University at: https://www.hse.ru/en/info/

[2] Implementation of some measures on openness of data on academic mobility of employees and their publications is observed in the universities participating in the so-called 5-100 project (Russian Academic Excellence Project 5-100. Read more at: https://www.5top100.ru/en/about/more-about/), but none of them had enough data for calculations in the framework of our approach and research design.

[3] This method is widely used in medical research as part of the design of an experiment to study the effectiveness of a particular treatment. In social research it is more often used in non-experimental design. The approach has been applied in studies of science, in particular in the analysis of the relationship between mobility and career achievements of scientists [Bäker 2015; Lawson, Shibayama 2015].

of long-term mobility and researchers with episodes of short-term mobility. As researchers with long-term mobility (long-mobile researchers) we determined those who on their personal pages or in CV marked at least one episode of abroad mobility with the length of one year and more. As a result, we have selected 78 long-mobile researchers (and 78 their non-mobile matched pairs) and 40 short-mobile researchers (and 40 their non-mobile matched pairs). For one researcher we could not detect the length of mobility.

Publication activity data for mobile researchers and their non-mobile matched pairs was taken from the largest international scientific citation database Scopus. If available, Scopus Author Identifiers (IDs) were derived from the personal pages of the studied sample of researchers; otherwise, we have run the search of Scopus author IDs. Further all the Scopus author IDs collected were checked on their correctness and completeness. Due to small size of our sample, author name disambiguation was performed in manual mode. Duplicate author profiles (cases when there are several Scopus author IDs for one author) were merged[4].

Further using the functionality of the Scopus SciVal analytical toolbox, we have created the synchronized corpus of Scopus author IDs for mobile and non-mobile researchers. In this database each row represents a specific mobile or non-mobile researcher from our sample with time series of some key bibliometric indicators like number of publications, number of citations received etc. for this specific researcher.

The preliminary analysis shows that first publications of mobile researchers appeared in 2004. We've further divided the 2004-2020 period onto three five-year sub-periods: 2005-2009, 2010-2014 and 2015-2019. We focus our analysis on 2015-2019 sub-period since during this period mobile and non-mobile researchers have published the highest number of Scopus-indexed publications.

We've integrated further all the publications of mobile and non-mobile researchers into two special corpuses of publications. In each corpus each row represents a single Scopus-indexed publication of mobile or non-mobile researcher with various data on this publication like a year of issue, a number of citation received, name of source issued, CiteScore value of this source etc. Aggregation of information from two corpuses of publications allowed us to calculate some basic bibliometric indicators (like number of publications, number of citations received, share of publication in international collaboration etc.) for two groups of researchers. All the calculations were performed for the Scopus data collected in October 2020.

To analyze the relationship between international mobility and the content of publications of mobile and non-mobile researchers we used data on so-called Scopus Topics Prominence in Science that is available in the SciVal analytical toolbox[5]. Among the various number of bibliometric indicators available in SciVal for each publication of mobile and non-mobile researchers we select for content analysis indicators like Topic Cluster name, Topic Cluster number, Topic name, Topic number, Topic Cluster Prominence Percentile, Topic Prominence. With these data, we can see for each researcher in what topics and clusters he or she has

[4] Merging of author profiles was done in the following cases: different versions of the transliteration of the surname and / or first name of a specific author; combining several author profiles with the same spelling of the surname, but different spelling options for the name and patronymic; combining profiles of the same author with different affiliations; unification of maiden surname and husband's surname for women. In some cases, instead of unification of duplicate profiles, we did the clearing of author profile (exclusion of publications that obviously do not belong to a specific author). Merging of Scopus Au-IDs in different formats and/or clearing of author profiles was done for about 40 cases of mobile and non-mobile researchers.

[5] Assigning of topic prominence in science to Scopus-indexed publications works as follows. To each specific publication indexed in Scopus special clustering algorithm automatically assigns one topic and one topic cluster (broader topic). Assigning of this or that topic to a specific publication is defined in the result of clustering based on direct citation analysis of almost all publications indexed in Scopus. As a result 96000+ topics and 1500+ topic clusters is defined for almost all Scopus-index publications. See more about topic prominence in Science at: https://www.elsevier.com/solutions/scival/releases/topic-prominence-in-science

published papers. The question we are interested in is there a difference in the size of "repertoire" of a researcher (the number of topics and the number of clusters) between mobile and non-mobile groups. Analysis of Scopus topics was done for the whole period: 2004-2020.

Results

Mobility and general research performance

Key results of bibliometric analysis of publication activity of short- and long-term mobile young researchers and their non-mobile matched pairs from HSE University in Scopus in 2015-2019 were summarized in Table 1.

Table 1. Key indicator of publication activity for the studied groups of mobile and non-mobile researchers from HSE University in Scopus for 2015-2019

Group of researchers and Type of mobility length → Indicator name ↓	Publications of mobile researchers			Publications of non-mobile matched pairs		
	All types	Long-term	Short-term	All types	Long-term	Short-term
Number of publications	724	475	250	595	398	185
Number of researchers who had at least one publication in Scopus	113/119	73/78	38/40	112/119	74/78	38/40
Average number of publications per one researcher of a group	6.08	6.09	6.25	5.00	5.10	4.63
Number of citations received	3 393	2 430	945	1 525	1 097	383
Average number of citations per one publications	4.67	5.12	3.78	2.56	2.76	2.07
Average value of field-weighted citation impact of all publications, points	0.96	1.03	0.83	0.70	0.72	0.62
Number of publications in Top 1 Citation Percentile (top-1% of the most cited publications)	3	3	0	0	0	0
Average value of CiteScore[1] of sources[2] where publications are issued, points	2.46	2.70	2.07	1.66	1.81	1.28
Share of publications in Q1 sources[1] by CiteScore (to publications in all sources with CiteScore calculated), %	35.3	37.5	32.3	25.2	26.4	20.8
Number of publications in top-1% sources[1] by CiteScore value	13	11	2	2	1	1
Share of publications in international collaboration, %[2]	42.0	46.5	34.4	19.8	21.3	17.3

Notes. 1. According to methodology of Scopus "Calculating the CiteScore is based on the number of citations to documents (articles, reviews, conference papers, book chapters, and data papers) by a journal over four years, divided by the number of the same document types indexed in Scopus and published in those same four years" – see more at: https://service.elsevier.com/app/answers/detail/a_id/14880/supporthub/scopus/. 2. Vast majority of sources for which CiteScore is calculated are journals, indexed in Scopus, and some conference proceedings, book series and books. 3. As publications in international collaboration we treat publications where the sole author or at least one co-author is affiliated with at least two different countries.

Source: authors' calculations based on Scopus SciVal data collected at October 2020. All types of documents indexed in Scopus are taken into account

Table 1 provides key bibliometric indicators for the corpuses of Scopus-indexed publications for the selected groups of researchers. As we can see from Table 1 almost all mobile researchers as well as their non-mobile matched pairs had at least one Scopus-indexed publication for 2015-2019. Our analysis shows that mobile researchers have in general better research performance indicators of Scopus-indexed publications than their non-mobile matched pairs. Mobile researchers outperform their non-mobile matched pairs by the average number of publications per one researcher.

Mobile researchers are especially stronger than their non-mobile matched pairs in terms of citation indicators, journal quality indicators and involvement in international research collaboration. Mobile researchers have much higher average number of citations per one publication, higher field-weighted citation impact of their publications, and publish their papers in journals with much higher average CiteScore metrics. Mobile researchers also have higher share of publications in Q1 journals and much higher number of publications in top-1% journals by CiteScore value. In addition, mobile researchers authored all publications in "top-1% most cited" category from the studied corpuses of publications. All these prove the first hypothesis (H1) that mobile researchers have better overall researcher performance.

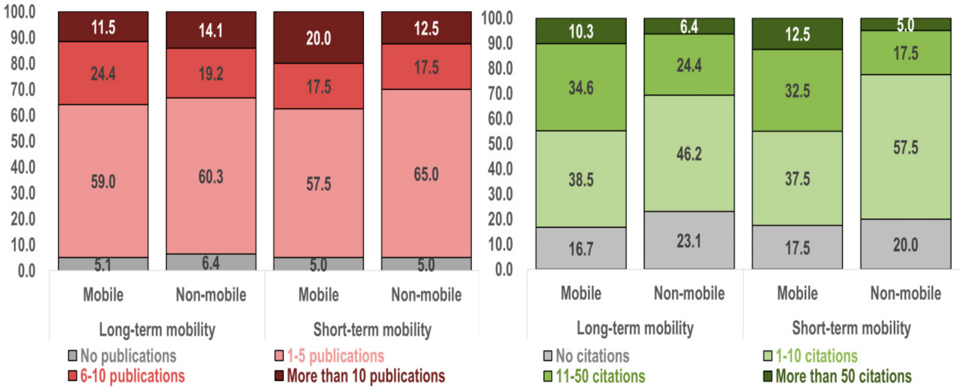
Higher average CiteScore value of journals where mobile researchers publish their work can be caused by a variety of factors. During their work or studying abroad mobile scientists can acquire new competencies that let them make better studies. Among these new competencies one can learn how to select research topics and questions to make the study eligible for publication in high-level international journal. Mobile researchers can also find new collaborators and join new projects producing high quality work. Apart from that mobile researchers could have some first-hand experience with editors or regular authors of high quality journals while working abroad. These social ties can also boost their chances to be published. Some of these possible explanations find evidence in bibliometric data we have.

Mobile researchers are much more actively involved in international research collaboration. It is not surprising that mobile researchers especially strongly outperform their non-mobile pairs by the level of integration into international research collaboration. After returning to Russia, mobile researchers bring their contacts with foreign co-authors, accumulated abroad, to their Russian places of work. The results of our analysis prove the second hypothesis (H2) about much higher collaboration of mobile researchers with foreign colleagues.

We found that that researchers who have in their career longer-than-one-year episode(-s) of international mobility in many aspects have better research performance indicators of their Scopus-indexed publications in comparison to researchers with short-term episodes of international mobility. Short-term mobile have slightly higher average number of publications per one researcher, but they lose in terms of involvement in international collaboration, average citation level and the average value of CiteScore indicator of journals where the publications were issued. We also see that vast majority of publications in top-1% sources by CiteScore value and all publications in Top 1 Citation Percentile are published by long-term mobile researchers. The conclusions that can be derived from the comparison of long-term mobile researchers with short-term mobile are rather limited though. These groups can differ in many aspects significant for publication performance. To analyze how the length of mobility is related to research performance we compared each group of mobile researcher to the appropriate control group of non-mobile researchers (matched pairs).

Differences between mobile and non-mobile researchers

Further we go to the level of individual authors to compare key indicators of publication activity of mobile researchers vs. their non-mobile pairs. Figure 1 shows the distribution of short- and long-term mobile researchers and their non-mobile matched pairs by the number of publications and number of citations received in Scopus for 2015-2019.



Source: authors' calculations based on Scopus SciVal data collected at October 2020. All types of documents indexed in Scopus are taken into account.

Figure 1. Distribution of researchers with long- and short-term mobility and their non-mobile matched pairs by number of publications and number of citation received in Scopus in 2015-2019

As we can see from this graph, the difference between mobile and non-mobile researchers by the level of publications activity is not dramatic. On the other hand, the differences by number of publications for mobile researchers vs. non-mobile pairs are stronger for the group with short-term mobility. Mobile researchers are in general slightly more productive than non-mobile researchers are. Among all groups studied short-term mobile researchers show the highest share of authors with 10+ publications in Scopus for 2015-2019. The differences in the number of citations received between mobile and non-mobile researchers are much stronger. Among mobile researchers, we can see a much less share of non-cited authors and much more share of authors who received 10 and more citations in comparison to non-mobile researchers. Meanwhile, there is no dramatic difference between short-term and long-term mobile researchers in distributions by number of citations received. Non-mobiles pairs of short-term mobile researchers show the "worst" distribution by number of citations received.

Table 2. Results of paired two-sample Student's t-test for mean number of publications and citations received for 2015-2019 in Scopus for mobile vs. non-mobile researchers

Type of distribution →	**Mean number of publications**			**Mean number of citations received**		
Type of mobility length → **Indicator name ↓**	**All types**	**Long-term**	**Short-term**	**All types**	**Long-term**	**Short-term**
N. of pairs	119	78	40	119	78	40
Mean value for mobile group	6.24	6.18	6.45	29.55	32.62	24.30
Mean value for non-mobile group	5.18	5.24	4.95	13.49	14.50	10.60
t-statistics	1.287	0.838	1.344	2.103	1.662	1.705
P-value (T <= t) one-sided	0.100	0.202	0.093	0.019	0.050	0.048
t critical one-sided	1.658	1.665	1.685	1.658	1.665	1.685

Notes: 1. Significance level (alpha-value) for all test is set as 5%. 2. Null hypothesis is that there are no differences in mean values of number of publications and citations received between mobile and non-mobile researchers. 3. Mean values of number of publications in Table 2 differs from Average number of publications per one researcher differs since some publications are coauthored by several from the studied samples.

Table 2 shows the results of the paired two-sample Student's t-test that was used to test the hypothesis about the difference of mean number of publications and citations received for 2015-2019 for two dependent samples of mobile vs. non-mobile researchers. The results of this test support our observations from Figure 1: differences between mobile and non-mobile researchers by number of citations received are stronger than differences by number of publications. The difference in number of publications between mobile and non-mobiles researchers proves statistically significant for short-term mobile group and all mobile, but not for the group of long-term mobile researchers. In case of citations received the difference between mobile and non-mobile is significant for all three groups (all, long-term, short-term).

Publication activity before and after international

The observed difference in bibliometric indicators for the two groups of researchers, even when proved statistically significant, does not automatically mean that it is an effect of international mobility. Despite the matched pairs selection procedure, the difference may be due to other factors that are not directly observable. For example, some personality traits could be the factor of both gaining foreign experience and publication activity.

Availability of data on the publications of researchers, and on the periods of their study and work abroad allows us to conduct additional analysis to partially exclude unaccounted factors. We compared the research performance of mobile and non-mobile researchers in periods before it could be affected by mobility. To do this analysis, for each mobile researcher we take into account only the papers published before the year of the first episode of mobility. For the non-mobile pair of this specific mobile researcher, we also take into account publications issued before the same year (Figure 2). As a result we have detected 102 mobile researchers (and 102 their non-mobile matched pairs) for whom we can detect the year of the first episode of international mobility.

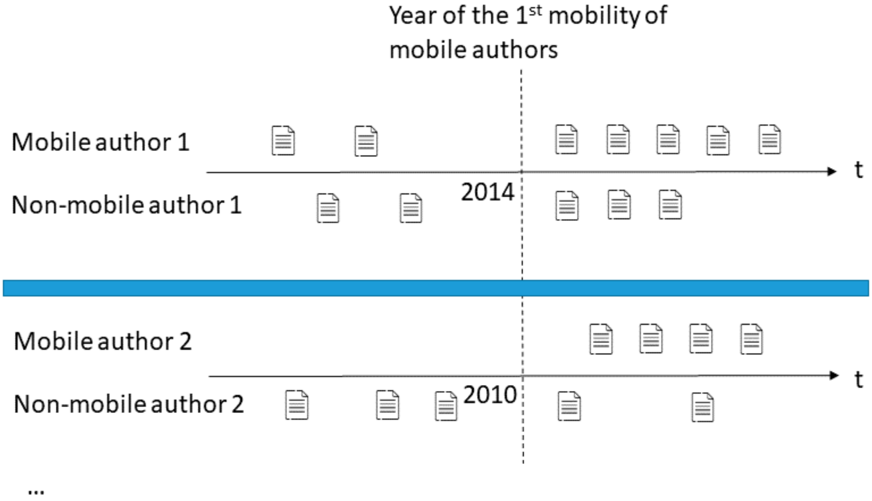


Figure 2. Publication activity of mobile researchers and their non-mobile matched pairs before and after the "treatment" (the first mobility).

Distribution of researchers according to number of papers published during two periods is shown on the Figure 3. On average, mobile researchers published 0.99 papers in Scopus before the first episode of international mobility, while the group of non-mobile researchers had published more intensively during the same periods – a mean is 1.44 publications. The median is 0 papers in both groups, and the difference of the means in not significant (p-value for a t-test is 0.268). As for publications in high-level journals, mobile researchers showed somewhat higher productivity at the start of a career – a mean is 0.22 papers in Q1 journals vs. 0.13 for non-mobile authors, but this difference is also not statistically significant. Thus, the advantage in publication activity of the group of mobile researchers was not noticeable at the start of their careers and appeared definitely after their work or studying abroad.

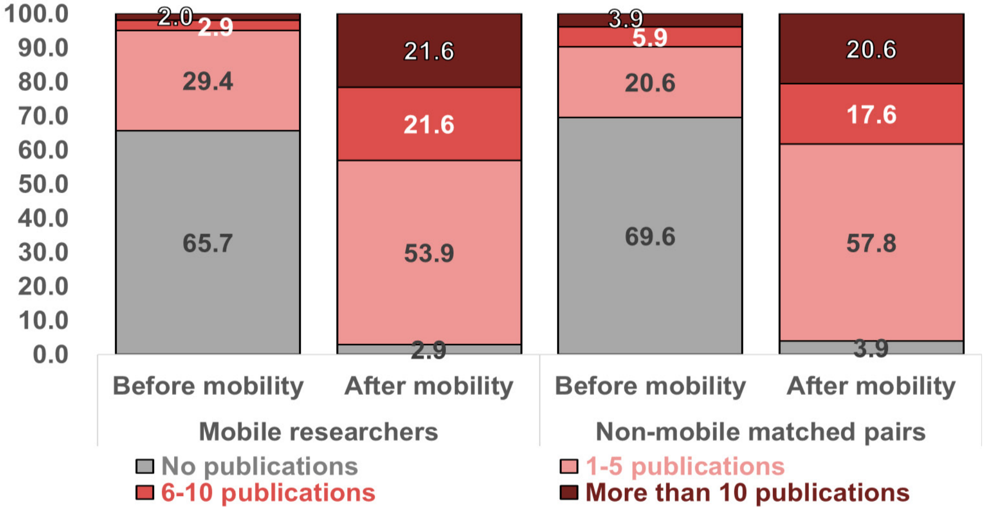


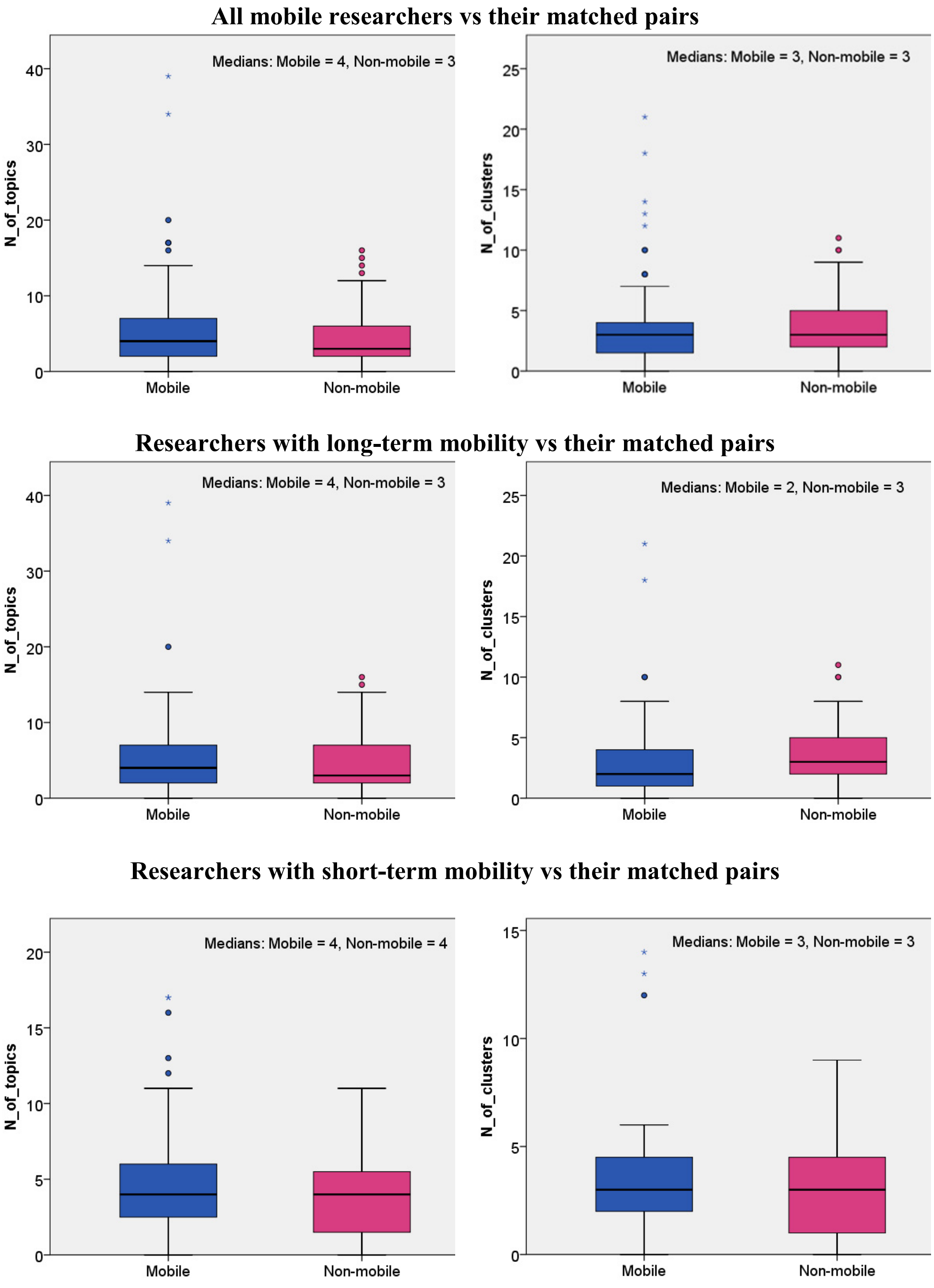
Figure 3. Distribution of mobile researchers and their non-mobile matched pairs according to number of papers in Scopus published before and after the year on the 1st mobility (and the same years for matched pairs).

Publication portfolios of mobile and non-mobile researchers

One of our goals was to investigate the difference between publication portfolios of mobile and non-mobile researchers beyond the traditional productivity and impact indicators. Particularly we are interested in research topics. One topic and one topic cluster are assigned to each publication in Scopus. Topic cluster is a group of closely related topics and can be regarded as a broad topic. Figure 4 shows the distribution of researchers in mobile and non-mobile groups by the number of topics and number of topic clusters in their papers.

We see that the distributions are not quite different. Mobile researchers tend to have more topics in their portfolio, than non-mobile. This is true for both long-term mobile researchers (mean number of topics is 5.3 vs 4.8 for non-mobile, medians are 4 vs 3) and short-term mobile ((means are 5.2 vs 3.8, medians are 4 vs 4)). The difference is statistically significant only for short-term mobile researchers (t-test, p-value < 0.09). When topic clusters (broad topics) are considered, there is no much difference in portfolio size of mobile and non-mobile researchers, it is not statistically significant at least. We can conclude that the hypothesis H3 is somewhat confirmed – mobile scientists make research on more diverse subjects than non-mobile – although the difference is not big, and is significant only in case of short-term mobility.

The difference observed does not mean that mobile researchers are actively engaged in research on more topics than non-mobile. On the contrary, the chances are that mobility actually helps researchers to focus, to find an "oilfield" and drill it. To investigate this in detail one need to look how the portfolio changes what a researcher moves to another country. We compared "active portfolios" of mobile and non-mobile researchers – the research topics in relatively recent publications (published in 2015-2019). Mobile group again turned out to have more diverse portfolios (mean number of topics 4.2 vs 3.4 for non-mobile, medians are 3 vs 3; t-test for means shows the difference is marginally significant with p-value =0.108).



All mobile researchers vs their matched pairs

Researchers with long-term mobility vs their matched pairs

Researchers with short-term mobility vs their matched pairs

Figure 4. Distribution of number of topics (left) and topic clusters (right) in papers published by mobile researchers and their matched pairs.

Discussion

In terms of basic indicators of activity, our results can be summarized as follows. Mobile researchers from HSE University have much better indicators of research performance in Scopus than researchers who have never worked or studied outside of Russia. Our findings support the similar others (Franzoni et al., 2015; Gibson & McKenzie, 2014; Gureyev et al., 2020; Netz, Hampel & Aman, 2020) and contradict to some others (Halevi et al. 2016; Cañibano et al. 2008). Moreover, continuing the discussion, opened by (Cañibano et al. 2020), we claimed that the period of mobility matters. Researchers with long episodes of international research or education mobility show even better research performance in Scopus.

In this paper, we open a direction in the analysis of publications of mobile researchers that seems promising to us – research topics analysis. We saw that researchers with mobility experience tend to make research in more topics than non-mobile researchers. The question is: "Whether it was the mobility that pushed the scientists to diversify their research interests?". One way to answer this question, which we hope to develop in future studies, is to analyze how the "repertoire" of researchers evolves over time – what are the most common patterns in research topics change associated with moving to another country. The evolution of the topic's portfolio can be related to a performance advantage, which is often registered for mobile researchers. It also contributes to the discussion about the pattern of change in research practices (Gläser et al. 2014).

We see the possible ways of development of our researcher as follows. One way – to analyze how the career stage when first episode of international mobility occurred matters for the further pattern of researcher performance, following the discussion by (Cañibano et al., 2020). The other way – is the analysis of disciplinary profiles of researchers (i.e. whether the difference between mobile researchers vs. non-mobile researchers is stronger in natural sciences vs. social sciences, following the work by (Laudel & Bielick, 2019).

Acknowledgments

This article was prepared with the support of the grant of the President of the Russian Federation for young scientists - Candidates of Sciences. Grant Number: MK-1418.2020.6.

References

Bäker A. (2015) Non-tenured post-doctoral researchers' job mobility and research output: An analysis of the role of research discipline, department size, and coauthors // *Research Policy*, vol. 44, no 3, pp. 634-650.

Cañibano C., Outamendi J., Andujar I. (2008) Measuring and assessing researcher mobility from CV analysis: the case of the Ramon y Cajal programme in Spain // *Research Evaluation*, vol. 17., no. 1, p. 17-31.

Cañibano C., D'Este, P., Otamendi, F. J., & Woolley, R. (2020). Scientific careers and the mobility of European researchers: an analysis of international mobility by career stage // *Higher Education*, vol. 80, no.6, p. 1175-1193.

De Filippo D., Sanz Casado E., Gomez I. (2009) Quantitative and Qualitative Approaches to the Study of Mobility and Scientific Perfomance: A Case Study of a Spanish University // *Research Evaluation*, vol. 18, no 3, pp. 191-200.

Edler J., Fier H., Grimpe C. (2011) International scientist mobility and the locus of knowledge and technology transfer // *Research Policy*, vol. 40, no 6, pp. 791-805.

Franzoni C., Scellato G., Stephan P. (2015) International mobility of research scientists: lessons from GlobSci. In A. Guena (Ed.) *Global Mobility of research scientist. The economics of who goes where and why* (pp. 35-65). Amsterdam, Elsevier. Chap. 2.

Gao Y., Liu J. (2020): Capitalising on academics' transnational experiences in the domestic research environment // *Journal of Higher Education Policy and Management*, doi:10.1080/1360080X.2020.1833276

Gibson J., McKenzie D. (2014) Scientific mobility and knowledge networks in high emigration countries: evidence from the Pacific // *Research Policy*, no. 43, p. 1486-1495.

Gläser J., Aljets E., Lettkehmann E., Laudel G. (2014) Where to go for a change? The impact of authority structures in universities and public research institutes on change of research practices // *Research in Sociology of Organizations*, vol. 42, pp. 297-329.

Gläser J., Laudel G. (2015) The three careers of an academic. *Zentrum Technic und Gesselshaft, TU Berlin, Discussion Paper* 35/2015.

Gureyev V.N., Mazov N.A., Kosyakov D.V., Guskov A.E. (2020) Review and analysis of publications on scientific mobility: assessment of influence, motivation, and trends // *Scientometrics*, vol. 124, pp. 1599–1630.

Halevi G., Moed H., Bar-Ilan J. (2016). Researchers' mobility, productivity, and impact: case of top producing authors in seven disciplines // *Publishing Research Quarterly*, no. 32, p. 22-37.

Horta, H., Jung, J., & Santos, J.M. (2019). Mobility and research performance of academics in city-based higher education systems // *Higher Education Policy*. doi:10.1057/s41307-019-00173-x

Jonkers K., Cruz-Castro L. (2013) Research upon return: The effect of international mobility on scientific ties, production and impact // *Research Policy*, vol. 42, no 8, pp. 1366-1377.

Jonkers, K., & Tijssen, R. (2008). Chinese researchers returning home: Impacts of international mobility on research collaboration and scientific productivity. *Scientometrics,* 77(2), 309-333.

Jöns, H. (2007). Transnational mobility and the spaces of knowledge production: a comparison of global patterns, motivations and collaborations in different academic fields. *Social Geography*, vol. 2, no. 2, 97-114.

Laudel, G., & Bielick, J. (2019) How do field-specific research practices affect mobility decisions of early career researchers? // *Research Polic*y, vol. 48, no.9, p. 103800.

Lawson C., Shibayama S. (2015) International research visits and careers: An analysis of bioscience academics in Japan // *Science and Public Policy*, vol. 42, no 5, pp. 690-710.

Netz N., Hampel S., Aman V. (2020) What effects does international mobility have on scientists' careers? A systematic review // *Research evaluation*, vol. 29, no 3, pp. 327-351.

Scellato G., Franzoni C., Stephan P. (2012) Mobile scientists and international networks // *National Bureau of Economic Research*. No. w18613.

Scellato G., Franzoni C., Stephan P. (2017) A mobility boost for research // *Science*, vol. 356, no 6339, pp. 694-697. DOI: 10.1126/science.aan4052

Shin, J., Jung, C., Postiglione, J., & Azman, G. (2014). Research productivity of returnees from study abroad in Korea, Hong Kong, and Malaysia // *Minerva*, vol. 52, no. 4, pp. 467–487. doi:10.1007/s11024-014-9259-9

Avoiding bias when inferring race using name-based approaches

Diego Kozlowski[1], Dakota S. Murray[2], Alexis Bell[3], Will Hulsey[3], Vincent Larivière[4], Thema Monroe-White[3] and Cassidy R. Sugimoto[2]

[1] *diego.kozlowski@uni.lu*
DRIVEN DTU-FSTM, University of Luxembourg, Luxembourg

[2] *dakmurra@iu.edu; sugimoto@indiana.edu*
School of Informatics, Computing, and Engineering, Indiana University Bloomington, IN, USA

[3]*alexis.bell@vikings.berry.edu; william.hulsey@vikings.berry.edu; tmonroewhite@berry.edu*
Campbell School of Business, Berry College, GA, USA

[4] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal, QC, Canada

Abstract
Racial disparity in academia is a widely acknowledged problem. The quantitative understanding of racial-based systemic inequalities is an important step towards a more equitable research system. However, few large-scale analyses have been performed on this topic, mostly because of the lack of robust race-disambiguation algorithms. Identifying author information does not generally include the author's race. Therefore, an algorithm needs to be employed, using known information about authors, i.e., their names, to infer their perceived race. Nevertheless, as any other algorithm, the process of racial inference can generate biases if it is not carefully considered. When the research is focused on the understanding of racial-based inequalities, such biases undermine the objectives of the investigation and may perpetuate inequities. The goal of this article is to assess the biases introduced by the different approaches used name-based racial inference. We use information from US census and mortgage applications to infer the race of US author names in the Web of Science. We estimate the effects of using given and family names, thresholds or continuous distributions, and imputation. Our results demonstrate that the validity of name-based inference varies by race and ethnicity and that threshold approaches underestimate Black authors and overestimate White authors. We conclude with recommendations to avoid potential biases. This article fills an important research gap that will allow more systematic and unbiased studies on racial disparity in science.

Introduction

The use of racial categories in the quantitative study of science is so widely extended that it intertwined with the controversial origins of statistical analysis itself (Galton, 1891; Godin, 2007). While Galton and the eugenics movement reinforced the racial stratification of society, racial categories have also been used to acknowledge and mitigate racial discrimination. As Zuberi (2001) explains: "The racialization of data is an artifact of both the struggles to preserve and to destroy racial stratification." This places the use of race as a statistical category in a precarious position: one that both reinforces the social processes that segregate and disempower parts of the population, while simultaneously providing an empirical basis for understanding and mitigating inequities.

Science is not immune from these inequities (Hoppe, et al., 2019; Prescod-Weinstein, 2020; Stevens et al, 2021; Ginther et al., 2011). Early research on racial disparities in scientific publishing relied primarily on self-reported data in surveys (e.g., Hopkins et al., 2013), geocoding (e.g., Fiscella & Fremont, 2006) and directories (e.g., Cook, 2014). However, there is an increasing use of large-scale inference of race based on names (Freeman & Huang, 2014), similar to the approaches used for gender-disambiguation (e.g., Lariviere et al., 2013). Algorithms, however, are known to encode human biases (Buolamwini & Gebru, 2018; Caliskan, 2017): there is no such thing as *algorithmic neutrality*. The automatic racial inferences of authors based on their features in bibliographic databases is itself an algorithmic process that needs to be scrutinized, as it can fall into implicitly encoded bias, with major impact in the over and under representation of racial groups.

In this study, we use the self-declared race/ethnicity from the 2010 U.S. census and mortgage applications as the basis for inferring race from author names on scientific publications indexed in the Web of Science (WoS) database. Bibliometric databases do not include self-declared race by authors, as they are based on the information provided in publications, such as given and family names. Given that the U.S. census provides the proportion of self-declared race by family name, this information can be used to infer US authors' race given their family names. We present several different approaches for inferring race and examine the bias generated in each case. The goal of the research is to provide an empirical critique of name-based race inference and recommendations for approaches that minimize bias. Even if a prefect inference is not achievable, the conclusions that arise from this study will allow researchers to conduct more careful analyses on racial and ethnic disparities in science. Although the categories analysed are only valid in the U.S. context, the general recommendation can be extended to any other country in which the Census (or similar data collection mechanism) includes self-reported race.

Racial categories in the US census

The U.S. census is a rich and long-running dataset, but also deeply flawed and criticized. Currently it is a decennial counting of all U.S. residents, whether citizens or non-citizens, in which several characteristics of the population are gathered, including self-declare race/ethnicity. The classification of race in the U.S. census is value-laden with the agendas and priorities of its creators, namely 18th century White men who Wilkerson (2020) refers to as "the dominant caste." The first U.S. census was conducted in 1790 and founded on the principles of racial stratification and White superiority. Categories included: "Free White males of 16 years and upward," "Free White males under 16 years;" "Free White females," "All other free persons," and "Slaves" (U.S. Bureau of the Census, 1975). At that time, each member of a household was classified into one of these five categories based on the observation of the census-taker, such that an individual of "mixed white and other parentage" was classified into "All other free persons" in order to preserve the "Free White…" privileged status. To date, anyone classifying themselves as other than "non-Hispanic White" is considered a "minority." The shared ground across the centuries of census survey design and classification strata reflects the sustained prioritization of the White male caste (D'Ignasio & Klein, 2020, Zuberi, 2001).

Today, self-identification is used to assign individuals to their respective race/ethnicity classifications (Locke, Blank, & Groves, 2011), per the U.S. Office of Management and Budget (OMB) guidelines. However, the concept of race and/or ethnicity remains poorly understood. For example, in 2000 the category "Some other race" was the third largest racial group, consisting primarily of individuals who in 2010 identified as Hispanic or Latino (which according to the 2010 census definition refers to a person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin regardless of race). Instructions and questions which facilitated the distinction between race and ethnicity began with the 2010 census which stated that "[f]or this census, Hispanic origins are not races" and to-date, in the U.S. federal statistical system, Hispanic origin is considered to be a separate concept from race. However, this did not preclude individuals from self-identifying their race as "Latino," "Mexican," "Puerto Rican," "Salvadoran," or other national origins or ethnicities (Humes, Jones, & Ramirez, 2011, p. 3). Furthermore, 6.1% of the US population changed their self-identification of both race and ethnicity between the 2000 and 2010 censuses (Liebler, et al., 2017), demonstrating the dynamicity of the classification. The inclusion of certain categories has also been the focus of considerable political debate. For example, the inclusion of citizenship generated significant debates in the preparation of the 2020 Census, as it can generate a larger nonresponse rate from the Hispanic community (Baum, 2019). The aim of this article, however, is to represent the full extent of possible US-based authors. Therefore, we consider both citizens and non-citizens.

In this paper, we design and compare multiple approaches for inferring race based on the given and family names of authors in the WoS database. The social function of the concept of race (i.e., the building of racialized groups) underpins its definition more than any physical traits of the population. For example, "Hispanic" as a category arises from this conceptualization, even though in the 2010 U.S. census the question about Hispanic origin is different from the one on self-perceived race. While Hispanic origin does not relate to any physical attribute, it is still considered a socially racialised group, and this is also how the aggregated data is presented by the Census Bureau. Therefore, in this paper, we will utilize the term race to refer to these social constructions, acknowledging the complex relation between conceptions of race and ethnicity. But even more important, this conceptualization of race also determines what can be done with the results of the proposed models. Given that race is a social construct, inferred racial categories should only be used in the study group-level social dynamics underlying these categories, and not as individual-level traits. Census classifications are founded upon the social construction of race and reality of racism in the US, which serves as “a multi-level and multi-dimensional system of dominant group oppression that scapegoats the race and/or ethnicity of one or more subordinate groups” (Horton & Sykes, 2001, p. 209). Self-identification of racial categories continue to reflect broader definitional challenges, along with issues of interpretation, and above all the amorphous power dynamics surrounding race, politics, and science in the U.S. In this study, we are keenly aware of these challenges, and our operationalization of race categories are shaped in part by these tensions.

Data

This project uses several data sources to test the different approaches for race inference based on the author's name. First, to test the interaction between given and family names distributions, we simulate a dataset that covers most of the possible combinations. Using a Dirichlet process (Teh, 2010), we randomly generate 500 multinomial distributions for given names, and another 500 random multinomial distributions for family names. After this, we build a grid of all the possible combinations of given and family names random distributions (250,000 combinations).

In addition to the simulation, we use two datasets with real given and family names and an assigned probability for each racial group. The data from the given names is from Tzioumis (2018), who builds a list of 4,250 given names based on mortgage applications, with self-reported race. Family name data is based on the 2010 US census (US Census Bureau, 2016), which includes all family names with more than 100 appearances in the census, with a total of 162,253 surnames that covers more than 90% of the population. For confidentiality, this list removes counts for those racial categories with fewer than five cases, as it would be possible to exactly identify individuals and their self-reported race. In those cases, we replace with zero and renormalize. As explained previously, changes were introduced in the 2010 US census racial categories. Questions now include both racial and ethnic origin, placing "Hispanic" outside the racial categories. The racial categories used in both datasets include Hispanic as a category, and all other racial categories excluding people with Hispanic origin, therefore the category "White" becomes "Non-Hispanic White Alone", and "Black or African American" becomes "Non-Hispanic Black or African American Alone", and so on. The final categories used in both datasets are:

- Non-Hispanic White Alone (*White*)
- Non-Hispanic Black or African American Alone (*Black*)
- Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander Alone (*Asian*)
- Non-Hispanic American Indian and Alaska Native Alone (*AIAN*)
- Non-Hispanic Two or More Races (*Two or more*)
- Hispanic or Latino origin (*Hispanic*)

We test these data on WoS to study how name-based racial inference performs on the population of U.S. scientific authors. WoS did not regularly provide first names in articles before 2008; therefore, the data includes all articles published between 2008 and 2019. This results in 21,295,333 articles, 1,609,107 unique first authors, 152,835 distinct given names and 288,663 distinct family names for first authors. Given that in this database, 'AIAN' and 'Two or more' account for only 0.69% and 1.76% of authors respectively, we remove these and renormalize the distribution with the remaining categories. Therefore, in what follows we will refer exclusively to categories *Asian, Black, Hispanic,* and *White*.

Methods

Manual validation

The data is presented as a series of distributions of names across race (Table 1). In name-based inference methods, it is not uncommon to use a threshold to create a categorical distinction: e.g., using a 90% threshold, one would assume that all instances of Juan as first name should be categorized as Hispanic and all instances of Washington as a given name should be categorized as Black. In such a situation, any name not reaching this threshold would be excluded (e.g., those with the last name of "Lee" would be removed from the analysis). This approach, however, assumes that the distinctiveness of names across races does not significantly differ.

Table 1. Sample of family names (US census) and given names (mortgage data).

Type	*Name*	*Asian*	*Black*	*Hispanic*	*White*	*Count*
Given	Juan	1.5%	0.5%	93.4%	4.5%	4,019
	Doris	3.4%	13.5%	6.3%	76.7%	1,332
	Andy	38.8%	1.6%	6.4%	53.2%	555
Family	Rodriguez	0.6%	0.5%	94.1%	4.8%	1,094,924
	Lee	43.8%	16.9%	2.0%	37.3%	693,023
	Washington	0.3%	91.6%	2.7%	5.4%	177,386

To test this, we began our analysis by manually validating name-based inference at three threshold ranges: 70-79%, 80-89%, and 90-100%. We sampled 300 authors from the WoS database, 25 randomly sampled for every combination of racial category and inference threshold. Two coders manually queried a search engine for the name and affiliation of each author and attempted to infer a racial category through visual perception and information listed on their websites and CVs (e.g., affiliation with racialized organizations such as *Black in AI, SACNAS* etc.).

Figure 1 shows the number of valid and invalid inferences, as well as those for whom a category could not be manually identified, and those for whom no information was found. Asian authors were valid at every threshold considered. The inference of Black authors, in contrast, was invalid or uncertain at the 70-80% threshold, but high at the 90% threshold. Similarly, inferring Hispanic authors was only accurate after the 80% threshold. Inference of White authors was valid at all thresholds but improved above 90%. This suggests that a simple threshold-based approach is not equally valid across all races. We thereby move to a weighting-based scheme for analysis that does not provide an exclusive categorization but uses the full information of the distribution.

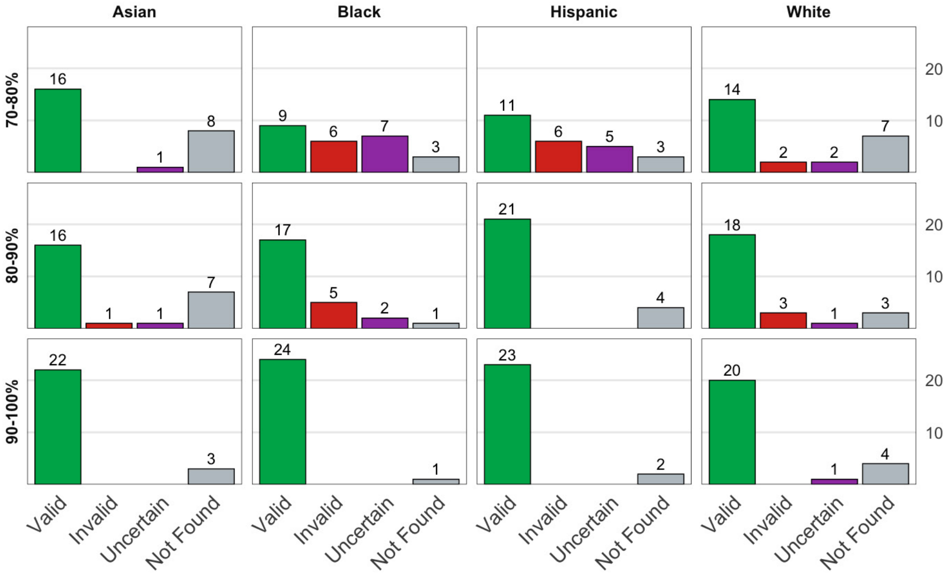


Figure 1. Manual validation of racial categories

Weighting scheme

We assess three strategies for inferring race from an author's name using a combination of their given and family name distributions across racial categories (Table 1). The first two aim to build a new distribution as a weighted average from both the given and family name racial distributions, and the third uses both distributions sequentially. In this section we explain these three approaches and compare them to alternatives that use only given or only family name racial distributions.

The weighting scheme should account for the intuition that if the given (family) name is highly informative while the family (given) name is not, the resulting average distribution should prioritize the information on the given (family) name distribution. For example, 94% of the people with Rodriguez as a family name identify themselves as Hispanic, whereas 39% of the people with the given name Andy identify as Asian, and 53% as White (see Table 1). For an author called Andy Rodriguez, we would like to build a distribution that encodes the informativeness of their family name, Rodriguez, rather than the relatively uninformative given name, Andy. The first weighting scheme proposed is based on the standard deviation of the distribution:

$$sd = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

With four racial categories, the standard deviation moves between 0, for perfect uniformity, and 0.5 when one category has a probability of 1. The second weighting scheme is based on entropy, a measure that is designed to capture the informativeness of a distribution:

$$e = -\sum_{i=1}^{c} P(x_i) log P(x_i)$$

Using this, we propose the following weight for given names:

$$x_{weight} = \frac{f(x)^{exp}}{f(x)^{exp} + f(y)^{exp}}$$

with x and y as the given and family name. exp is the exponent applied to the function, and is a tuneable parameter. For the standard deviation, using the square function means we use the variance of the distribution. In general, the higher the exp is set, the more skewed the weighting is towards the most informative name distribution.

Figure 2 shows the weighting to the given and family names based on their informativeness, and for different values of the exponent. The horizontal and vertical axes show the highest value on the given and family name distribution, respectively. This means that a higher value on any axis corresponds with a more informative given/family name. The color palette shows how much weight is given to the given names. When the exponent is set to two, both the entropy and standard deviation-based models skew towards the most informative feature, a desirable property. Compared to other models, the variance gives the most extreme values to cases where only one name is informative, whereas the entropy-based model is the most uniform.

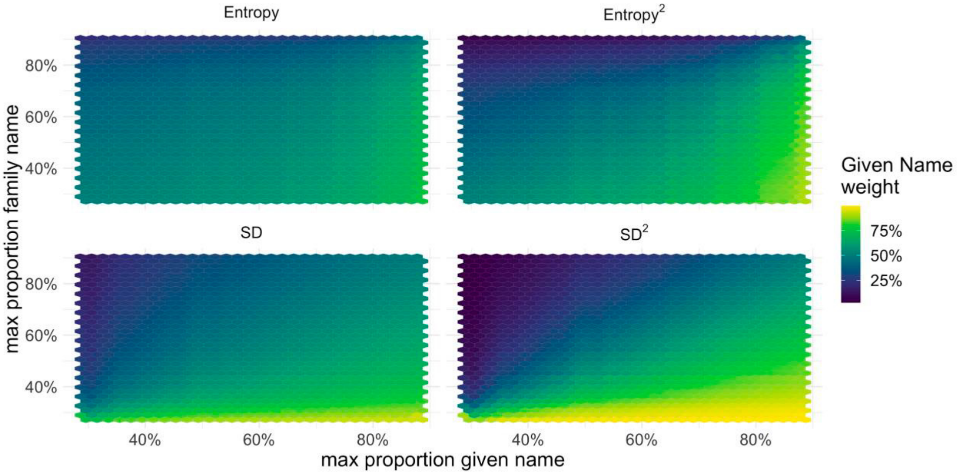


Figure 2. Given names weight distribution by given and family name skewness. Simulated data.

Information retrieval

The above weighting schemes result in a single probability distribution of an author belonging to each of the racial categories, from which a race can be inferred. One strategy for inferring race from this distribution is to select the racial category above a certain threshold, if any. A second strategy is to make use of the full distribution to weight the author across different racial categories, rather than assigning any specific category. Third, we implement a two-step strategy that sequentially uses family and then given names. Given that family names are from census data, we first retrieve all authors which have a family name with a probability of belonging to a specific racial group greater than a given threshold. This retrieves *N* authors. Second, we retrieve the same amount, *N,* of authors as in the first step, but using the given names. Finally, we define a unique set of authors, removing those captured by both steps. This leaves a set between *N* and 2*N* authors, depending on the number of authors retrieved in both steps. Naturally, there are multiple possible changes on this two-step method, for example, a percentage threshold could be used in both steps, instead of a given number. It could be possible also to have a specific percentage threshold for each step. There is not a preferred method in principle, but the proposed configuration relates with our belief that the census data (family names) is of better quality, and therefore we prefer to use it to define the number observations to extract with the lower quality data.

Results

The effect of underlying skewness

Before comparing the results of the proposed strategies for using both given and family names, we present characteristics of these two distributions on the real data, and in relation to the WoS dataset. Table 2 shows the population distribution on the family names, based on the US census, and on the given names, based on the mortgage applications. Considering the US census data as ground truth, we see that the mortgage data highly over-represents the White population, particularly over-represents Asians, and underrepresents Black and Hispanic populations; this likely stems from the structural factors (i.e., economic inequality, redlining, etc.) that prevent marginalized groups from applying for mortgages in the US. People may also choose to self-report a different racial category when responding anonymously to the census bureau than when applying for a mortgage loan. Due to this bias in the distribution of given names, we normalize the given name distribution by applying an expansion factor to each racial group, so that the expanded given name count preserves the racial distribution in the US census.

Table 2. Racial representation of family names (US census) and given names (mortgage data).

Racial group	*Family names*	*Given names*
White	66.1%	82.6%
Hispanic	16.5%	6.9%
Black	12.4%	4.2%
Asian	5.0%	6.3%

Both given and family names share a characteristic not considered in our simulated data: the informativeness of names varies across racial groups. Inferring racial categories based on a set threshold will, then, produce biased results as typical names of one racial category are more informative, and thus more easily meet the threshold, than another. Figure 3 shows how the representation of inferred races changes based on the assignment threshold used. Increasing the threshold results in fewer total individuals returned (top), as some names are not very sufficiently informative. For family names, only a small proportion of the population remains at a 90% threshold. The Asian population is highly over-represented between the 90% and 96% threshold, after which they suddenly become under-represented. The White population is systematically over-represented for any threshold, whereas the Black population is systematically underrepresented. The Hispanic population is overrepresented between the 65% and 92% threshold and underrepresented after. With given names, the White population is systematically overestimated for every threshold until 96%, where the Asian population is suddenly overestimated to a high degree. Similar results are observed based on given names. Again, the Asian population is highly overrepresented after the 96% threshold, whereas the White population is over-represented across nearly all thresholds and the Black and Hispanic population were underrepresented across all thresholds. The fact that Asian, and to some degree Hispanic, populations have more informative given and family names is related with higher differentiation of these two communities; White and Black populations in the United States, in contrast, tend to have more similar names (as verified in Elliott et al., 2009). Given that the White population is larger than the Black population, the use of a threshold (and assigning all people with that name to a single category), generates a Type I error on Black authors, and Type II error on White authors, thereby overestimating the proportion of White authors.

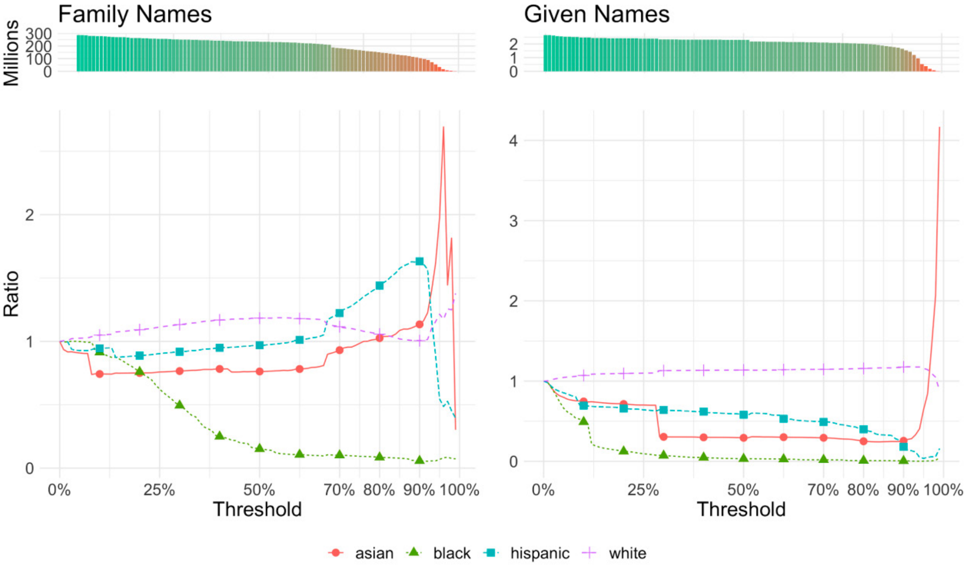


Figure 3. Changes in groups share, and people retrieved, by threshold. Census (Family names) and mortgage (Given names) datasets.

The effect of thresholding

Figure 4 shows the effect of using a 90% threshold on the WoS dataset of unique authors. The first column corresponds to each author counting fractionally towards each racial category in proportion to the probabilities of their name distribution, using family names from the census, i.e., this is the closest we can get to ground truths with the available information. The remaining columns represent (1) inference based on given and family names alone; (2) combination distributions; and (3) the two-steps strategy, using both normalized and unnormalized given names, always with a 90% threshold. All models highly underrepresented the Black population of authors. All models except normalized given names under-represent the Hispanic population. The unnormalized given name, either alone or in the variance model, underrepresents the Asian population. Finally, the White population is overrepresented by all models except family names and the variance with normalized given names.

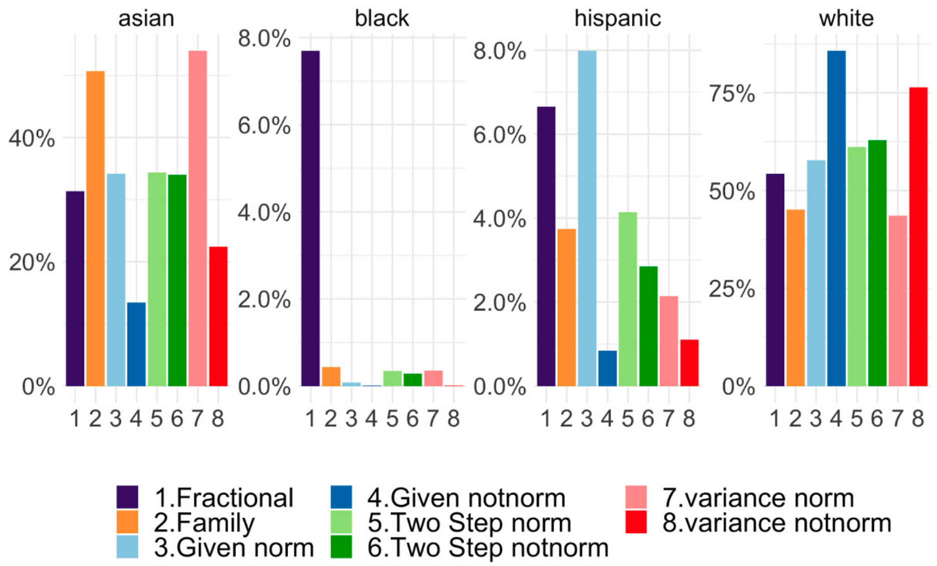


Figure 4. Resulting distribution on different models with 90% threshold. Fractional counting on family names for comparison.

Figure 5 demonstrates how the number of authors in each racial group drops when increasing the threshold. The convexity in the Black authors curve, distinct from the concave curves for other racial groups, explains the underrepresentation of this group under high thresholds. Even at less extreme thresholds, the proportion of Black authors in the dataset is far smaller than the fractionalized counting of White authors (851,843.60). The consequence of this is that a low threshold would grossly over represent White authors, while a high threshold would exclude nearly all Black authors. We conclude from this that a threshold-based approach, while intuitive and straightforward, should not be used for racial inference. Rather, analysis should be adapted to consider each author as a distribution over every racial category; in this way, even though an individual cannot be assigned into a category, aggregate results will remain unbiased.

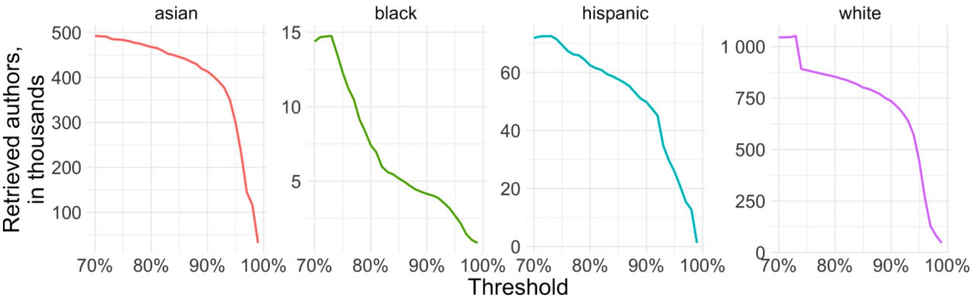


Figure 5. Two-steps strategy retrieval.

The effect of imputation

Another consideration is how to deal with unknown names. As mentioned in the Data section, the family names dataset provided by the census bureau covers 90% of the US population. The remaining 10%, as well as author names not represented in the census, generates 774,381 articles, or 18.75% of the dataset, for which the family name of the first authors has an unknown distribution over racial categories.

An intuitive solution would be to impute missing names with a default distribution based on the racial composition of the entire census. Table 3 shows the distribution among racial groups in the US census and in WoS (for first authors with family names included in the US census data). The Asian population is highly overrepresented among WoS authors, whereas Hispanic and Black authors are highly underrepresented, with respect to the total US population. Imputing with the census-wide racial distribution is, therefore, equivalent to skewing the distribution towards Hispanic and Black authors and underrepresenting Asian authors. Since the ground truth is contingent to the specific dataset in use, the natural imputation would instead be the mean of the population most representative of an individual. For example, in the case of a missing author name in the WoS, the racial distribution of that individual's discipline could be imputed. Our recommendation is, therefore, to first compute the aggregate distribution of racial categories with the dataset in which the inference is intended, and then use this aggregate distribution to impute in those family names missing from the census dataset. Statistically, this preserves the aggregate distribution on this dataset.

Table 3. Racial distribution in US census and WOS US Authors with known family names.

Racial group	*US census*	*US WoS*
White	66.1%	59.4%
Hispanic	16.5%	5.4%
Black	12.4%	7.2%
Asian	5.0%	24.5%

Conclusion

Race scholars (Emirbayer & Desmond, 2011) have advocated for a renewal of Bourdieu's (2001) call for reflexivity in science of science. We pursue this through empirical reflexivity: challenging the instrumentation used to collect and code data for large-scale race analysis. In this paper we manually validate and propose several approaches for name-based racial inference of US authors. We demonstrated the behaviour of the different methods on simulated data, across the population, and on authors in the WoS database. We showed the risks of underestimating highly minoritized groups (e.g., Black authors) in the data when using a threshold, and the overestimation of White authors introduced by given names when they are based on mortgage data. A similar result was identified by Cook (2014), in her attempt to infer race of patent data based on the US census: she found that the approach "significantly underpredicted matches to black inventors and overpredicted matches to white inventors" and concludes that the name-based inference approach was not suitable for historical analyses. From our analysis, we come away with three major lessons that are generally applicable to the use of name-based inference of race in the US, shown in table 4.

Inferring race based on name is an imperfect, but often necessary approach to studying inequities and prejudice in bibliometric data (e.g., Freeman & Huang, 2014), and in other areas where self-reported race is not provided. However, the lessons shown here demonstrate that care must be taken when making such inferences in order to avoid bias in our datasets and studies.

Table 4. General recommendations for implementing a name-based inference of race.

	Do's	*Don'ts*
Given Names	Use only family names from US census to avoid bias.	Do not use given names, except the underlying distribution of your dataset matches that of mortgage data.
Thresholding	Consider each person in your data as a distribution and adapt your summary statistics.	Do not use a threshold for categorical classification of each person, as this underrepresents Black population, due to the correlation between racial groups and name informativeness.
Imputation	Calculate first the aggregated distribution on your dataset, and use this for imputation of missing cases.	Do not use the census aggregate distribution for imputation, except your target population matches the US population.

It has been argued that science and technology serve as regressive factors in the economy, by reinforcing and exacerbating inequality (Bozeman, 2020). As Bozeman (2020) argued, "it is time to rethink the economic equation justifying government support for science not just in terms of why and how much, but also in terms of who." Studies of the scientific workforce that examine race are essential for identifying who is contributing to science and how those contributions change the portfolio of what is known. To do this at scale requires algorithmic approaches; however, using biased instruments to study bias often replicates the very inequities they hope to address.

In this study, we attempt to problematize the use of race from a methodological and variable operationalization perspective in the US context. However, any extension of this work across country lines will necessarily require tailoring to meet the unique contextual needs of the country or region in question. Ultimately, scientometric researchers utilizing race data are responsible for preserving the integrity of their inferences by situating their interpretations within the broader socio-historical context of the people, place, and publications under investigation. In this way, they can avoid preserving unequal systems of race stratification and instead contribute to the rigorous examination of race and science intersections toward a better understanding of the science of science as a discipline. As Zuberi (2001) remarked: "The racialization of data is an artifact of both the struggles to preserve and to destroy racial stratification."

References

Bozeman, B. (2020). Public Value Science. *Issues in Science and Technology*, 34-41.

Bourdieu, P. (2001). *Science of Science and Reflexivity.* Chicago, IL: University of Chicago Press.

Buolamwini, J. & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Caliskan, A., Bryson, J. J. & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186.

Cook, L.D. (2014). Violence and economic activity: evidence from African American patents, 1870-1940. Journal of Economic Growth, 19, 221-257.

D'Ignazio, C. & Klein, L. F. (2020). *Data feminism*. MIT Press.

Elliott, M.N., Morrison, P.A., Fremont, A., McCaffrey, D.F., Pantoja, P. & Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9, 69.

Emirbayer, M. & Desmond, M. (2011). Race and reflexivity. *Ethnic and Racial Studies*, 35(4), 574-599.

Fiscella, K. & Fremont, A.M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research*, 41(1), 1482-1500.

Freeman, R.B. & Huang, W. (2014). Collaborating with people like me: Ethnic co-authorship within the US. NBER working paper 19905.

Galton, F. (1891). *Hereditary genius*. D. Appleton.

Ginther, D.K., Basner, J., Jensen, U., Schnell, J., Kington, R. & Schaffer, W.T. (2018). Publications as predictors of racial and ethnic differences in NIH research awards. *PLoS ONE*, 13(11), e0205929.

Ginther, D.K., Schaffer, W.T., Schnell, J., Masimore, B., Liu, F., Haak, L.L. & Kington, R. (2011). Race, ethnicity, and NIH research awards. *Science*, 333(6045), 1015-1019.

Godin, B. (2007). From eugenics to scientometrics: Galton, Cattell, and men of science. *Social Studies of Science*, 37(5), 691-728.

Hopkins, A.L., Jawitz, J.W., McCarty, C., Goldman, A. & Basu, N.B. (2013). Disparities in publication patterns by gender, race and ethnicity based on a survey of a random sample of authors. *Scientometrics*, 96, 515-534.

Hoppe, T.A., Litovitz, A., Willis, K.A., Meseroll, R.A., Perkins, M.J., Hutchins, B.A., Davis, A.F., Lauer, M.S., Valantine, H.A., Anderson, J.M. & Santangelo, G.M. (2019). Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. *Science Advances*, 5: eea7238.

Horton, H. D. (1998). Toward a critical demography of race and ethnicity: Introduction of the "R" word. *Sociology Faculty Scholarship*, 1. https://scholarsarchive.library.albany.edu/sociology_fac_scholar/1

Horton, H. D. & Sykes, L. L. (2001). Reconsidering wealth, status, and power: Critical Demography and the measurement of racism. *Race and Society,* 4(2), 207-217.

Horton, H. D. (2002). Rethinking American diversity: Conceptual and theoretical challenges for racial and ethnic demography. In N. Denton & S. Tolnay (Eds.), American diversity: A Demographic challenge for the twenty-first century (p. 261–278). New York: State University of New York Press.

Humes, K., Jones, N. & Ramirez, R. (2011). Overview of Race and Hispanic Origin: 2010. U.S. https://10.4:51awww.atlantic.org/images/publications/Democratic_Defense_Against_Disinformation_FINAL.pdf

Larivière, V., Ni, C., Gingras, Y., Cronin, B. & Sugimoto, C.R. (2013). Global gender disparities in science. *Nature*, 504, 211-213.

Liebler, C.A., Porter, S.R., Fernandez, L.E., Noon, J.M. & Ennis, S.R. (2017). America's Churning Races: Race and Ethnicity Response Changes Between Census 2000 and the 2010 Census. *Demograph*y, 54(1), 259-284.

Locke, G., Blank, R. & Groves, R. (2011). 2010 Census Redistricting Data (Public Law 94-171) Summary File. https://www.census.gov/prod/cen2010/doc/pl94-171.pdf

Prescod-Weinstein, C. (2020). Making Black women scientists under white empiricism: the racialization of epistemology in physics. *Signs: Journal of Women in Culture and Society*, *45*(2), 421-447.

Stevens, K. R., Masters, K. S., Imoukhuede, P. I., Haynes, K. A., Setton, L. A., Cosgriff-Hernandez, E., ... & Eniola-Adefeso, O. (2021). Fund Black scientists. *Cell*, 184(3), 561-565.

Teh, Y. W. (2010). Dirichlet Process. https://www.stats.ox.ac.uk/~teh/research/npbayes/Teh2010a.pdf

Tzioumis, K. (2018). Demographic aspects of first names. *Scientific data*, 5, 180025.

U.S. Bureau of the Census. (1975). Historical Statistics of the United States, Colonial Times to 1970, Bicentennial Edition, Part 1). https://www.census.gov/history/pdf/histstats-colonial-1970.pdf

US Census Bureau. (2016). Frequently Occurring Surnames from the 2010 Census. The United States Census Bureau. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

Wilkerson, I. (2020). *Caste: The Origins of Our Discontents*. Random House.

Zuberi, T. (2001). *Thicker than blood: How racial statistics lie.* Univ. of Minnesota Press.

Links from preprints to published papers in preprint metadata

Bianca Kramer[1]

[1] *b.m.r.kramer@uu.nl*
Utrecht University Library, Utrecht University, Heidelberglaan 3, 3584 CX Utrecht (The Netherlands)

Abstract

Preprints have become an important part of the scholarly communication ecosystem. To be able to connect preprints to subsequent journal publications, and thereby have access to the full record of versions of a publication, the existence of reliable links between preprints and subsequent publications, available in an open infrastructure, is important. This paper reports on research in progress investigating, for a subset of COVID19-related preprints, authoritative links between preprints and published papers in Crossref metadata and on preprint servers themselves. It is shown that the coverage of links from preprints to published papers in Crossref metadata is often incomplete compared to links found on preprint servers themselves, underlining the potential for improvement in the update of metadata.

Introduction

In a growing number of scientific disciplines, preprints have become an important part of the way research results are communicated. As an example, over the last year, COVID19-related preprints have been shared on over 35 different preprint servers (Fraser and Kramer, 2020). Some of these are disciplinary preprint servers, most often on a non-profit basis (e.g. bioRxiv, medRxiv (both hosted by Cold Spring Harbor Laboratory), SocArXiv and PsyArXiv (both hosted on the Open Science Framework (OSF)). Other preprint servers are associated with legacy publishers, either directly linked to their submission workflow (e.g. ResearchSquare used by Springer Nature, and JMIR Preprints by JMIR) or also open to preprints independent of submission to the publisher's own journals (e.g. Preprints.org from MDPI, SSRN from Elsevier). Yet another example is ChemRxiv, a disciplinary preprint server backed by a number of scholarly societies including the American Chemical Society (ACS) and the Royal Society for Chemistry (RSC).

In order for preprints to form an integral part of the scholarly record, it is important to be able to link them to subsequent journal publications. This will enable access to the full record of versions of a publication (e.g. to track changes over time), irrespective of where each version is published. Many preprint servers (including all the examples mentioned above) use Crossref to obtain DOIs for their preprints, and consequently, register preprint metadata with Crossref. Crossref notifies preprint servers of potential matches with published articles. Crossref notifies preprint servers of potential matches with published articles. It requires preprint servers to verify the links and add them to the metadata record of the preprint. (Crossref, 2020). In addition, preprint servers also often add links to published papers on the landing pages of preprints.

Open availability of authoritative links from preprints to published papers in a centralized infrastructure (such as with Crossref) as open metadata, without restrictions on use and reuse, makes this information available for other systems to integrate and build upon, e.g. in discovery systems, for evaluation purposes and for transparent analysis on developments in scholarly communication. As such, they contribute to making research publications not only accessible and reusable, but also findable and interoperable (Waltman, 2020).

This research-in-progress investigates, for a subset of COVID19-related preprints, authoritative links between preprints and published papers in Crossref metadata and on preprint servers themselves. Furthermore, it uses these data to look at time between publishing of the preprint and publishing of the published paper for different preprint servers, as well as the destination (at the level of publisher) of published preprints, as these can both influence the observed links to published papers in preprint metadata.

Methods

Corpus of COVID-19 related preprints

As corpus for this investigation, COVID-19 related preprints with Crossref DOIs were used, as collected by Fraser and Kramer (2020). This corpus was collected by querying the Crossref REST API for all records with publication type 'posted content' and posted date between January 1, 2020 and April 11, 2021 indicated. Preprints were subsequently classified as being related to COVID-19 on the basis of keyword matches in their titles or abstracts (where available). The search string was defined as: coronavirus OR covid-19 OR sars-cov OR ncov-2019 OR 2019-ncov OR hcov-19 OR sars-2. Preprints were deduplication within preprint servers (keeping only the earliest posted version) but not between different preprint servers.

Links between preprints and published papers in Crossref

For all COVID-19 related preprints in the corpus defined above, information was collected on links to published papers, by querying the Crossref REST API for all DOIs and retrieving the information in the metadata field 'metadata field relation.is-preprint-of', which contains the DOI of the published paper. Subsequently, information on the published paper (journal, publisher and date of publication) was collected via a separate query on the Crossref REST API.

Links between preprints and published papers on bioRxiv and medRxiv

For COVID-19 related preprints in the corpus defined above that were published on bioRxiv and medRxiv, links to published papers were collected by querying the biorXiv API for all DOIs in our corpus of preprints, based on code used in Fraser et al., 2021. Subsequently, information on the published papers (journal, publisher and date of publication) was collected via a separate query on the Crossref REST API.

Data collection, analysis and data visualization

Following data collection, Data on links between preprints and published papers, time between preprint publication and journal publication and destination of published preprints were analyzed and visualized. Full R scripts for data collection, analysis and visualization are available on Github (Kramer, 2021). All data were collected on April 25, 2021.

Results

Links to published papers in preprint metadata.

Overall, the rate of COVID19-related preprints with links to published papers in Crossref metadata is only 11% (4146 of 36541 preprints with Crossref DOI). A number of preprint servers do not include links to published papers in their metadata on Crossref, including SRRN (n=5862 COVID19-related preprints in this dataset), Authorea (n=1356), and Scielo Preprints (n=312).

Among preprint servers that do include links to published papers in their metadata, the proportion of preprints linked to published papers, ranges from 7.7% (for OSF) to 51.3% (for JMIR) (Figures 1,2).

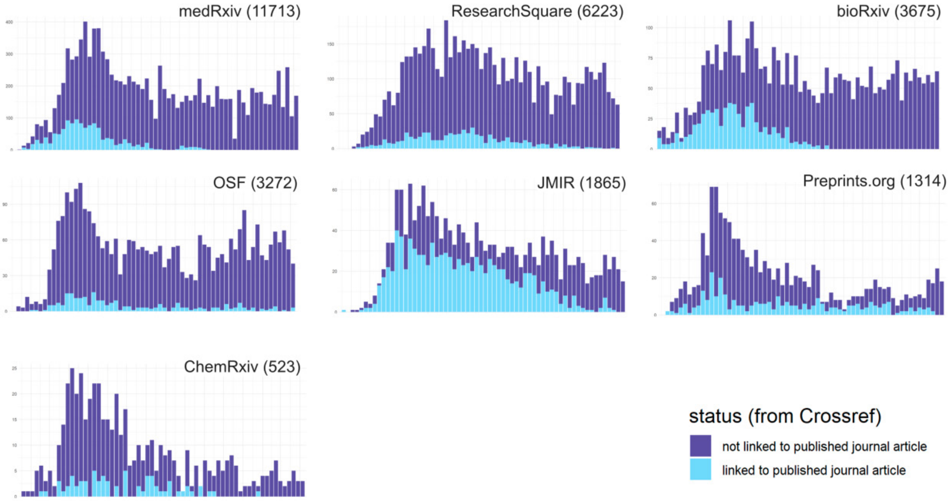


Figure 1. COVID-19 related preprints per week (January 2020-April 2021) with and without links to published papers in Crossref metadata.

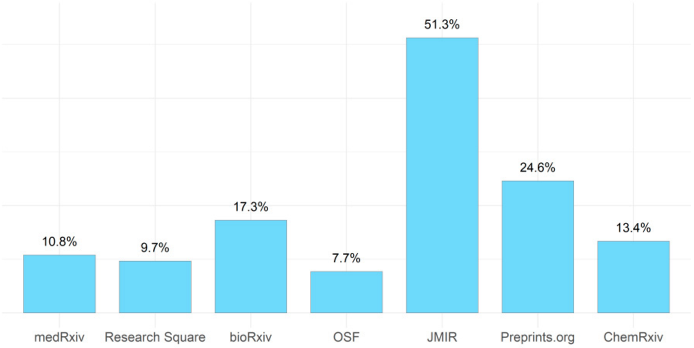


Figure 2. Percentage of COVID-19 related preprints (January 2020-April 2021) with links to published papers in Crossref metadata.

Time to publication

For preprints in this sample with a link to a subsequent journal article, the average time to publication (measured as the difference between the posted date of the first version of the preprint and the publication date of the subsequent journal article in Crossref metadata) is 97 days (close to 3 months). There is no clear difference between preprint servers in time to publication - preprint servers with a relatively high proportion of preprints with a link to a published paper (esp. JMIR) do no have a shorter average time to publication (Figure 3). OSF shows the largest spread in time to publication, which could be due to the variety of preprint servers using the OSF platform, with corresponding differences in publication cultures including (timing of) preprint sharing. Time to publication can also be negative, reflecting cases where the preprint is shared after publication of the journal article.

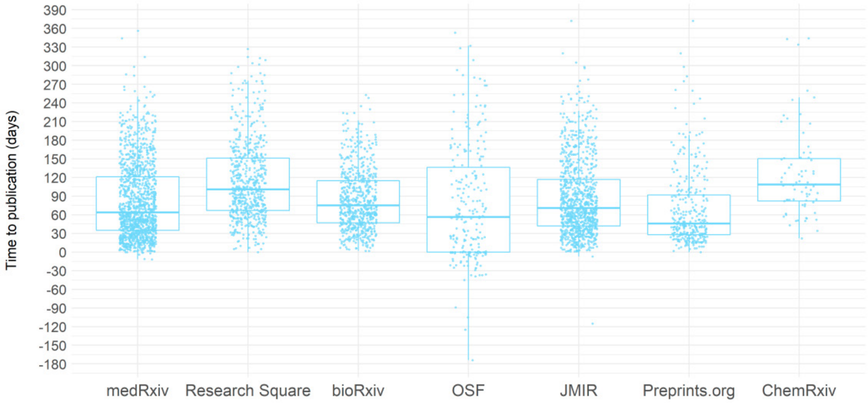


Figure 3. Distribution of time to publication (in days) for COVID-19 related preprints with links to published papers in Crossref metadata, for different preprint servers.

Links to published papers on bioRxiv and medRxiv

Both bioRxiv and medRxiv have more extensive coverage of published articles on their platform itself than recorded in their preprints' metadata: 28.7% vs. 10.5% for medRxiv and 32.7% vs. 17.3% for bioRxiv, for COVID19-related preprints in this sample (Figures 4, 5). NB. There were no cases of preprints with only a link to a published paper in the metadata, but not on the preprint platform.

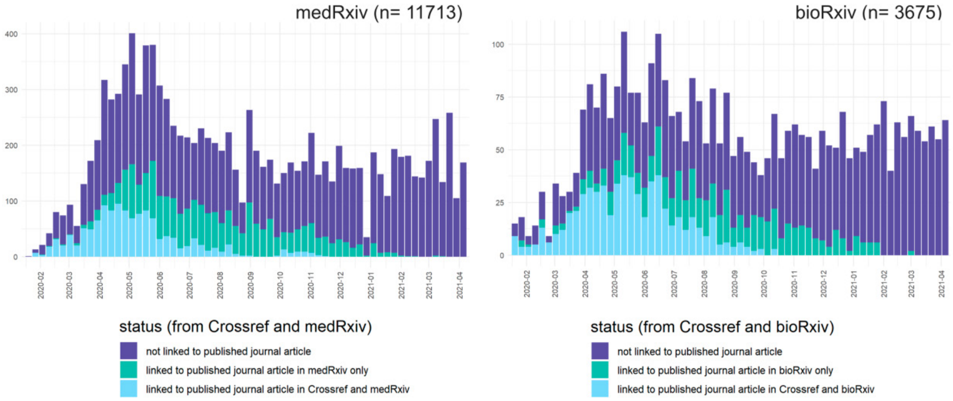


Figure 4. COVID-19 related preprints per week (January 2020-April 2021) on medRxiv and bioRxiv with links to published papers in Crossref metadata, on the preprint platform only, or neither

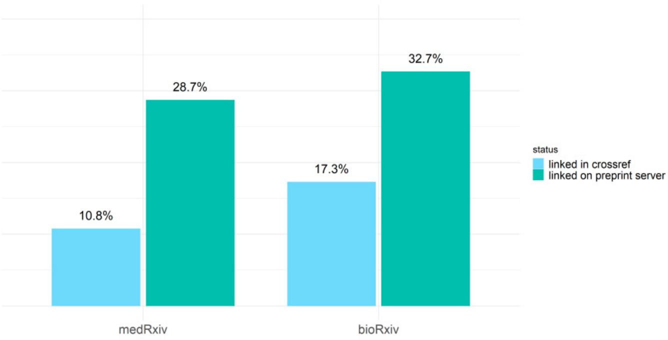


Figure 5. Percentage of COVID-19 related preprints (January 2020-April 2021) on medRxiv and bioRxiv with links to published papers in Crossref metadata or on the preprint platform.

Destination of published preprints

An alluvial plot was made showing the destination of all preprints with links to a published paper in their metadata. As expected, preprints from publisher-associated preprint servers JMIR and ResearchSquare predominantly are published in journals from JMIR and SpringerNature, respectively. However, only a subset of preprints on Preprints.org with a link to a subsequent paper get published in MDPI-journals, with over half being published in journals from other publishers.

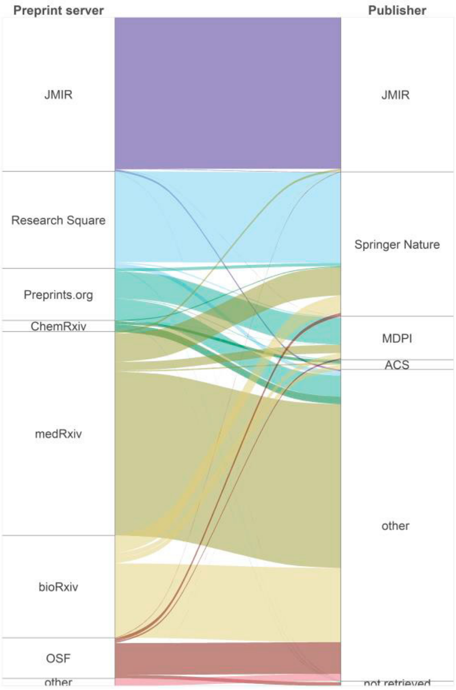


Figure 6. Destination of COVID-19 related preprints (January 2020-April 2021) with links to published papers in Crossref metadata.

Discussion

Coverage of links to published articles in preprint metadata in Crossref is expected to be incomplete. Not all preprint servers include such links in their metadata, and those that do might do so with a time delay and matches might be missed. Among preprint servers that do include links to published papers in their metadata, differences in the proportion of preprints linked to published papers, could reflect both technical workflows (e.g. linking might be easier/quicker when preprint server and journals are from the same publisher) and publication practices (e.g. selectivity of journals, speed of peer review processes, decisions on when to post a preprint).

There appears to be no clear difference between preprint servers in time to publication - preprint servers with a relatively high proportion of preprints with a link to a published paper do not have a shorter average time to publication. It might also be expected that linking preprints and published papers might be easier/quicker when preprint server and journals are associated with the same publisher, and indeed, JMIR, and to a lesser extent Preprints.org and ResearchSquare, have the highest proportion of preprints linked to published papers in the sample studied here.

Both bioRxiv and medRxiv have more extensive coverage of published articles on their platform itself than recorded in their preprints' metadata. The delay in updating this information in metadata records points to the potential for more accurate and complete coverage of links to published papers in metadata of preprints.

Having authoritative links from preprints to published papers available as open metadata will benefit the scholarly communication system. It will also be interesting to investigate the potential of additional similarity-based matching of preprints to published papers (see e.g. Lachapelle 2020, Cabanac et al., 2021), such as in EuropePMC (that links preprints and published papers), Unpaywall (that includes preprints as green open access versions of published papers) and Microsoft Academic (that groups detected versions of a paper in a 'paper family').

Acknowledgments

The author thanks Nicholas Fraser and Jeroen Bosman for helpful feedback and discussions on this project. This submission is made available under a CC-BY 4.0 license.

References

Cabanac, G. et al. (2021). Day-to-day discovery of preprint–publication links. *Scientometrics*, 10.1007/s11192-021-03900-7.

Crossref (2020). Posted content (includes preprints) markup guide. Retrieved February 14, 2021 from: https://www.crossref.org/education/content-registration/content-type-markup-guide/posted-content-includes-preprints/.

Fraser, N. & Kramer, B.M.R. (2020). COVID-19 Preprints. *Github*. Retrieved January 21, 2021 from: https://github.com/nicholasmfraser/covid19_preprints.

Fraser, N. et al. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLoS Biol,* 19(4), e3000959.

Kramer, B.M.R. (2021). COVID-19 Preprints. *Github*. Retrieved February 14, 2021 from: https://github.com/bmkramer/covid19_preprints_published.

Lachapelle, F. (2020). COVID-19 Preprints and Their Publishing Rate: An Improved Method. *medRxiv,* 2020.09.04.20188771

Waltman, L. (2020). Publications should be FAIR. *Leiden Madtrics.* Retrieved February 14, 2021 from: https://leidenmadtrics.nl/articles/publications-should-be-fair.

Do impact-factor journals make blacklisted journals visible? A systematic citation study

Emanuel Kulczycki[1*], Marek Hołowiecki[1], Zehra Taşkın[1] and Franciszek Krawczyk[1]

[1] *Scholarly Communication Research Group, Adam Mickiewicz University in Poznań (Poland)*
* emek@amu.edu.pl

Abstract

Predatory publishing has recently become the most popular and controversial subject in academia. The main reason is the efforts of policymakers to draw a line between predatory and legitimate publishing. While the journals included in the citation indexes are accepted as 'legitimate' by policymakers, the journals in two lists (Beall's and Cabell's), one of which has not been updated for years, are considered blacklisted. Our main aim in this study is to reveal the contribution of the journals accepted as legitimate by policymakers to the visibility of predatory journals. The research in progress presents preliminary findings regarding the first stage of a two-part study. In the first stage, the contribution of impact-factor journals to the visibility of predatory journals will be revealed, and in the second, citations will be examined in terms of their contents. In this study, the indexes (JCR, ESCI or A&HCI) and impact factors of the citing journals are presented. According to the preliminary results of the study, the results confirm the necessity of citation content analysis.

Introduction

Predatory publishing is one of the most discussed topics regarding journal publishing in academia, which crosses over narrow fields of bibliometrics, scientometrics and academic-publishing studies (Frandsen, 2017; Xia et al., 2015). This topic related to publishing in so-called questionable or low-quality journals attracts much attention not only in academia but also outside it (Bohannon, 2013; Sorokowski et al., 2017). Predatory journals, accused of damaging science and diminishing the quality of scholarly communication and trust in science, are trying to be classified and listed. In recent years, the most famous attempt to list predatory journals was initiated by Jeffrey Beall, whose list (henceforth: Beall's list) gained attention from scientists and the media (Krawczyk & Kulczycki, 2020). The second well-known approach is done by the company Cabell's International (henceforth: Cabell's list), which not only lists predatory journals but also offers another product listing 'reputable' journals. Thus, journals listed on Beall's or Cabell's lists are called in this study blacklisted journals in contrast to whitelisted journals (listed for instance in reputable international indexes like Scopus or Web of Science Core Collection).

This study is the first extensive study looking at citations from Web-of-Science-indexed journals to papers published in blacklisted (or predatory) journals. Demir (2018) pointed out the big difference between the largest citation databases: Scopus indexes 53 predatory journals from Beall's lists, but Web of Science indexes only three such journals, and Somoza-Fernández et al. (2016) reported that this difference is small but still visible. This could also suggest a difference in citations to predatory journals in these databases, but previous attempts at analysing such citations were based mostly on Google Scholar data (due to an easier data-acquisition procedure) or Scopus.

In this paper, we will provide a systematic-citation study focussing on journals covered by the WoS products and on citers in terms of their country affiliation. The study aim is two-fold: The first is to analyse the visibility of blacklisted journals, and the second is to understand citation contexts. This paper is the first paper of a bigger endeavour aiming to investigate how papers published in blacklisted journals from the field of social sciences are cited by papers published by journals indexed in the Web of Science Core Collection. This paper will be followed by an upcoming study on the content-based analysis of citations of the blacklisted papers to evaluate the citations in terms of their content (meaning, purpose, shape, array). This is the key reason why we focus on social-sciences journals here: evaluating the content of citations will require expertise in the field (Cano, 1989).

Data and Methods

We have used two blacklists: Beall's and Cabell's lists. Moreover, we have collected journals' ISSNs from their websites and used the ISSN Portal to find variants of journal titles and ISSNs as well as to provide data of the country of publishing. The data on citations of papers published in selected journals were obtained from the Web of Science Core Collection (WoS) using the Cited Reference Search. We focussed on three main WOS products: Journal Citation Reports (JCR) based on the Science Citation Index Expanded and the Social Sciences Citation Index, the Arts and Humanities Citation Index (A&HCI) and the Emerging Sources Citation Index (ESCI). We downloaded PDF files of either paper from blacklisted journals and WoS-indexed journals to collect and verify the country affiliations of the authors. Using the blacklisted journal websites, we collected the data on the number of papers published by each journal.

We have decided to include in the analysis only active journals that have fulfilled the set of criteria: they published at least one paper in each year of the 2012–2018 period, their websites were active at the moment of the start of this study (May 2019), were not indexed in WoS even only for one year in at least one of three selected WoS products during the years 2012–2018, and were classified (on expert-scope decisions of the authors of this study) as social-sciences journals according to their titles and aims and scopes published on journals' websites. In the final sample, we have 74 unique blacklisted journals (37 indexed in Beall's and 37 indexed in Cabell's). Ten of those journals are indexed on both lists.

We have prepared two datasets to analyse citations. The first one consists of the bibliographic data and PDF files of the papers published in the years 2012–2018 in the social-sciences blacklisted journals included in this study and were cited by journals indexed in WoS (henceforth: cited papers). The other dataset consists of data on the papers (i.e., all journal publication types) published in journals indexed in WoS in the years 2012–2019 (henceforth: citing papers) that cited papers from the first dataset. The articles from 9 of 74 journals were not cited by articles from WoS-indexed journals, or their PDF files were missing on journals' websites. This allows us to analyse 3,234 unique cited papers from 65 blacklisted journals and 5,964 unique citing papers (6,750 citations of cited papers) from 2,338 WoS journals. The list of analysed blacklisted journals is in Appendix.

Results

Share of cited papers from blacklisted journals

Sixty-five analysed journals published 25,146 papers between 2012 and 2018, of which 3,234 (13%) were cited by WoS journals.

Figure 1 shows the highest share, i.e., 19%, in 2012 and 2013. The mean number of papers published by a single journal in the analysed period is 53.5 (min=43, max=2,176). On average, 11% of papers published by a journal were cited by WoS journals. The highest shares were found for journals that published 1,748 and 259 papers. The shares are 36% (635 papers) and 35% (91 papers), respectively. In the analysed sample, 8,327 papers were published in journals listed in Beall's list (10.3% of them were cited) and 13,910 papers from journals indexed in Cabell's list (14.7% were cited). 5,587 papers were published by journals indexed in both Beall's and Cabell's lists.

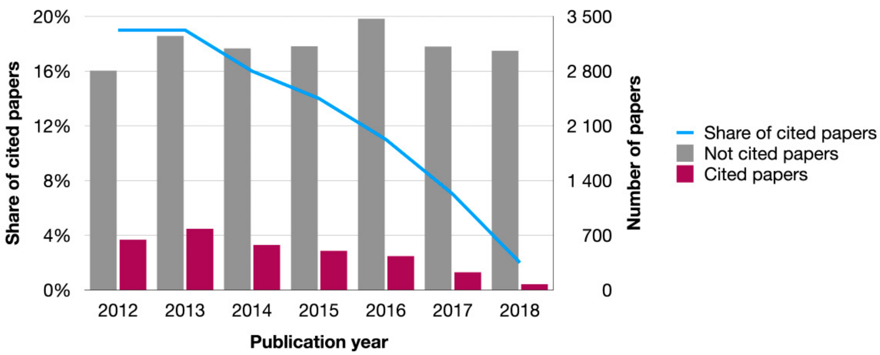


Figure 1. Share and Number of Cited Papers from Blacklisted Journals by WoS Journals

We checked whether 65 analysed journals were covered by Scopus, as this could potentially influence the share of citations. We found that five of 65 were or have been covered: one has been covered by the whole analysed period (24% of papers from this journal were cited), one removed from Scopus before the analysed period (23%), one covered and removed in the analysed period (36% papers), one covered in the last year of the analysed period (19%) and one covered before the analysed period and removed during the period (8%).

Web of Science journals citing blacklisted journals

We found that 2,338 unique WoS journals cited 3,234 blacklisted papers 6,750 times. The average number of citations per blacklisted journal from WoS journals is 2.88 (median=1, minimum=1, maximum=218). Half of the citations were from 261 journals. Eighty-nine of 2,338 journals cited papers from blacklisted journals at least 10 times, and four WoS journals cited over 100 blacklisted papers. In the analysed period, one WoS journal published 183 papers, which cited blacklisted papers from our sample 218 times (all except one published in one blacklisted journal). One of these blacklisted papers was cited 36 times by this WoS journal. We analysed in which WoS product (JCR, ESCI, A&HCI) a journal was indexed when a citing paper was published. We consider the publication year and whether a journal was included in a WoS product in the year in question. We found that 1,152 of 2,338 journals were indexed in ESCI, 35 in A&HCI and 1,047 in JCR. One-hundred and four journals that published 366 citing papers were neither in ESCI, A&HCI nor in JCR, which means that they were either in SCIE or in SSCI indexes but not yet JCR (e.g., waiting for calculation of their impact factor) or dropped from the indexes because of quality issues or manipulations such as citation stacking or excessive self-citation rates. Figure 2 shows how blacklisted journals were cited by WoS journals. Of the 6,750 citations, 2,502 (37%) were from JCR journals, 3,821 (56.6%) were from ESCI journals and 61 (0.9%) were from A&HCI journals.

Impact-factor journals citing blacklisted journals

The fact that a journal has a valid impact factor or is included in citation indexes is used by policymakers and managers to determine the level of that journal. We analysed the relationship between the impact-factor journals and their citations to blacklisted journals. To be able to make accurate statistical analyses, the impact factor of all journals cited in blacklisted journals were gathered with yearly changes. For example, if two articles in the same journal cited the blacklisted journals in 2018 and 2019, JCR 2017 and JCR 2018 were used. As a result, 1,600 impact factors for 1,047 IF journals were obtained.

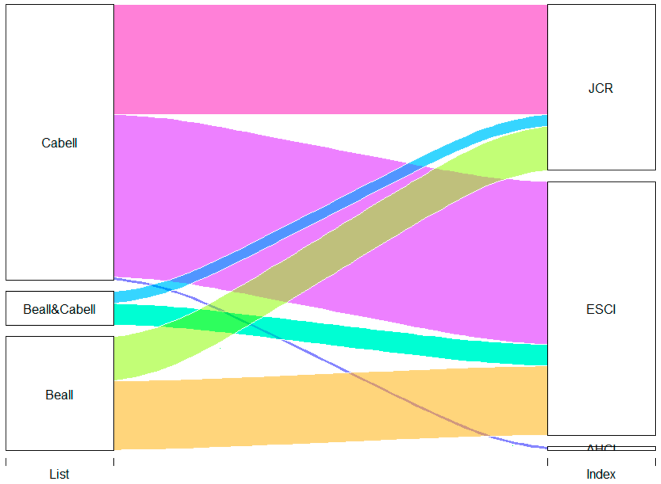


Figure 2. WoS Journals Citing Blacklisted Journals According to Blacklists

Before presenting the comparisons between impact factors and citations to blacklisted journals, it is worth mentioning those journals dropped from citation indexes. Twenty impact-factor journals that cited blacklisted journals 125 times were dropped from JCR or WoS for different reasons. Fifteen of them were dropped from the index without listing any unethical concerns. This means that coverage of the journals did not meet the WoS selection criteria (Clarivate, 2018). *Scientific World Journal* was suppressed from JCR based on citation stacking and four journals (*Business Ethics: A European Review*, *Environmental Engineering and Management Journal, Eurasia Journal of Mathematics Science and Technology Education* and *Industria Textila*) were dropped for their excessive self-citation rates. These five journals cited blacklisted journals 39 times. Furthermore, although they were not indexed in JCR and did not have an impact factor, 15 journals were excluded from ESCI after being indexed for a couple of years in ESCI. All these findings can be commented as questionable journals in WoS cited blacklisted journals; however, statistical tests did not confirm this comment.

The Spearman's Rho correlation coefficient shows that the correlation between journal impact factors and the number of citations to blacklisted journals is very low, at a 99% confidence level (r=0.090, p<0.001). Also, according to the Kruskal-Wallis test results, the differences between journal impact-factor quartiles and the number of citations to blacklisted journals were not significant (χ^2= 7.785, df=3, p=0.051). However, the Mann Whitney U test revealed that the only differences were found between Q1 and Q4 journals' number of citations to blacklisted journals (U=72661.500, Z=-2.648, p=0.008).

The impact-factor range of blacklisted journal citers is from 0 to 27.604 (mean=1.689, median=1.378, SD=1.471, 25%=0.745, 75%=2.252), while the minimum impact factor of the whole JCR between 2011 and 2018 is 0 and the maximum is 244.585 (mean=2.072, median=1.373, SD=3.310, 25%=0.704, 75%=2.462).

Eighty percent of the journals in JCR cited blacklisted journals only one time, and there is a significant difference between the impact factors of one-time citers and the others (U=174977.000, Z=-3.668, p<0.001). However, the surprising result is that the average impact factor of journals that cited blacklisted journals more than once is 1.896 (median=1.634), and this is higher than that of one-time citers (mean=1.639, median=1.318).

Table 1 shows the main features of 10 impact-factor journals that cited blacklisted journals more than 20 times.

Table 1. Ten Impact-Factor Journals that Cited Blacklisted Journals more than 20 Times

Journal name	*IF 2019*	*N of articles citing blacklisted journals*	*N of citations to blacklisted journals*	*Publisher country*	*Journal self-citation rate of the journal*
ACTA MEDICA MEDITERR	0.249	25	126	Italy	63.7%
SUSTAINABILITY-BASEL	2.576	56	56	Switzerland	38.9%
SAGE OPEN	0.715	16	45	USA	2.8%
SYSTEM	1.979	31	34	England	12.4%
FRONT PSYCHOL	2.067	24	26	Switzerland	15.5%
EGIT BILIM	0.493	19	25	Turkey	14.0%
PLOS ONE	2.740	25	25	USA	5.1%
COMPUT EDUC	5.296	21	22	England	11.6%
INT J BANK MARK	2.800	11	21	England	33.7%
COMPUT HUM BEHAV	5.003	20	21	USA	11.4%

All the test results on impact-factor journals prove that it is impossible to evaluate the blacklisted journals by looking at the impact factors or impact-factor percentiles of the journals because no pattern is identified. The impact factor is neither a descriptor of the quality of a paper nor the quality of citation. Therefore, it reveals the importance of content-based analysis in understanding the purpose of citations to blacklisted journals.

Discussion and Conclusions

The main aim of our study is to reveal the contributions of citation indexes, which are accepted as the authority in research evaluations, to the visibility of blacklisted journals, whose scientific levels are always considered quite low in academia. According to the results, 13% of the blacklisted articles were cited by Web of Science journals and 37% of these citations came from the impact-factor journals. If we accept being cited from authority-citation indexes as a tool for visibility, it is obvious that the indexes help the blacklisted journals to be visible regardless of the name of the index, whether SSCI, A&HCI or ESCI. The question to be asked at this point is: Do citations to the blacklisted journals make citation indexes questionable, or do these citations require a closer look at articles published in blacklisted journals? It is easy to accept all the papers published in blacklisted or questionable journals as low-quality, but without answering the question, it is difficult to draw a boundary for the definition between high and low quality. The findings show that there are no significant differences between the impact factors and the number of citations to the blacklisted journals. All the statistical tests conducted using journal-level statistics in our study confirm the emerging need to analyse the citations to blacklisted journals at the article level.

This is the first study in which a large-scale analysis of citations to predatory journals is conducted using WoS. When compared to the different results present in the citation studies based on Scopus, it is difficult to assess differences in terms of citing predatory journals in these two databases. However, taking into account that only 13% of articles in our study are cited, we can be sure that citations to the predatory journal are much more frequent in Google Scholar because it was reported there that 43% of articles were cited (Björk et al., 2020). Since we did not assess the quality of cited papers published in blacklisted journals, there are two possible interpretations of the main result of our study: 1) up to 13% of worthless articles in predatory journals can still leak to the mainstream literature legitimised by WoS, and 2) up to 13% of papers published in blacklisted journals are somehow important for developing scholarly legitimate discussion in social science. Unlike Oermann et al. (2020), we are not so sure that the important conclusion of the studies on predatory journals is to stop citing them completely. We prefer to leave the question raised by the result of our study open.

Our results indicate that there is no connection between the value of JIF of a given journal and this journal's citations to predatory journals. Although the number of such citations is relatively small, it could be another argument against treating JIF as a measure of journals' quality.

Acknowledgements

The work was financially supported by the National Science Centre in Poland (Grant Number UMO-2017/26/E/HS2/00019).

Appendix

The list of 65 analysed blacklisted journals is available here:
https://figshare.com/articles/dataset/Appendix_–_List_of_65_blacklisted_journals/13560326

References

Björk, B.-C., Kanto-Karvonen, S., & Harviainen, J. T. (2020). How Frequently Are Articles in Predatory Open Access Journals Cited. *Publications*, *8*(2), 17. https://doi.org/10.3390/publications8020017

Bohannon, J. (2013). Who's Afraid of Peer Review? *Science*, *342*(6154), 60–65. https://doi.org/10.1126/science.342.6154.60

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, *40*(4), 284–290.

Clarivate. (2018, June 27). *Web of Science: Editorial statement about dropped journals*. https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Editorial-statement-about-dropped-journals?language=en_US

Demir, S. B. (2018). Scholarly databases under scrutiny. *Journal of Librarianship and Information Science*, 096100061878415. https://doi.org/10.1177/0961000618784159

Frandsen, T. F. (2017). Are predatory journals undermining the credibility of science? A bibliometric analysis of citers. *Scientometrics*, *113*(3), 1513–1528. https://doi.org/10.1007/s11192-017-2520-x

Krawczyk, F., & Kulczycki, E. (2020). How is open access accused of being predatory? The impact of Beall's lists of predatory journals on academic publishing. *The Journal of Academic Librarianship*, 102271. https://doi.org/10.1016/j.acalib.2020.102271

Oermann, M. H., Nicoll, L. H., Ashton, K. S., Edie, A. H., Amarasekara, S., Chinn, P. L., Carter-Templeton, H., & Ledbetter, L. S. (2020). Analysis of Citation Patterns and Impact of Predatory Sources in the Nursing Literature. *Journal of Nursing Scholarship*, *52*(3), 311–319. https://doi.org/10.1111/jnu.12557

Somoza-Fernández, M., Rodríguez-Gairín, J.-M., & Urbano, C. (2016). Presence of alleged predatory journals in bibliographic databases: Analysis of Beall's list. *El Profesional de La Información*, *25*(5), 730. https://doi.org/10.3145/epi.2016.sep.03

Sorokowski, P., Kulczycki, E., Sorokowska, A., & Pisanski, K. (2017). Predatory journals recruit fake editor. *Nature*, *543*(7646), 481–483. https://doi.org/10.1038/543481a

Xia, J., Harmon, J. L., Connolly, K. G., Donnelly, R. M., Anderson, M. R., & Howard, H. A. (2015). Who publishes in "predatory" journals? *Journal of the Association for Information Science and Technology*, *66*(7), 1406–1417. https://doi.org/10.1002/asi.23265

Constructing models of academic funding system: relation between the amount of funding and the efficiency

Lai Kwun Hang[1], Vincent Traag[2] and Ludo Waltman[3]

[1] *k.h.lai@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (the Netherlands)

[1] *V.A.Traag@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (the Netherlands)

[1] *waltmanlr@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (the Netherlands)

Abstract

The academic funding system has been under criticism for many years. While there are some on-going experiments trying to improve the system, there isn't any clear conclusion or guideline. On the other hand, modeling is a cost-effective way to facilitate the discussion. In this paper, we present a construction of analytical and simulation models to study the relation between the amount of funding and the efficiency of the funding system. In many of the settings, we show that there could exist some optimal amounts of funding to maximize the efficiency. The results are useful to reflect on the funding system in the real world.

Introduction

Issues in the academic funding system are attracting a lot of attention in the recent years. As an essential part of academia, the funding system is one of the major factors contributing to over-competitiveness in academia (Carson et al., 2013). Though a certain level of competitions may improve the resource allocation process, over-competitiveness may do more harm than good. Reports claim that scholars are pressured to seek out external funding, or even to change their focus of research to increase their chance of funding success (Roberts, 2007). As a result, rather than doing actual research, researchers may dedicate a large amount of effort to writing funding applications.

Several proposals were made to address this problem in the funding system. A well-known example is the introduction of a lottery for awarding funding (Fang & Casadevall, 2016. The Health Research Council of New Zealand has run an experiment on implementing a lottery system (Adam,2019). While some people are positive about the approach, some researchers oppose such an allocation method. Most researchers who applied to a funding program with a lottery element reported that the lottery did not decrease their preparation time (Liu et al., 2020). These experiments are useful to test different possibilities for implementing alternative funding allocation modes. However, running experiments is quite expensive and time-consuming, limiting the number of alternatives that are explored. In contrast, modeling and simulations provide a low-cost and flexible methodology for exploring alternatives. Some earlier studies have used analytical models (Gross & Bergstrom, 2019) or agent-based models (Avin, 2015; Geard & Noble, 2010) to model different aspects of the academic funding system.

We are interested in better understanding the efficiency of the funding system. In particular, our central question is how the efficiency changes with the amount of funding. To attempt to answer this question, we develop and present several analytical models. We provide analytical results, but for more complex cases we provide results based on extensive simulations. We here construct models of the funding system, starting from the simplest case and gradually adding more realistic elements. Our results suggest that under some conditions increasing the amount of funding may decrease the efficiency by increasing the competition too much.

Although our models only capture a small portion of the possible dynamics, having a systematic construction may improve our understanding of the results. Additionally, our approach may enable us to relate our models to previous studies and study the differences. Finally, it may facilitate extending and modifying the models to study other closely related problems.

Models

We model the funding system in a utility-based framework. Scholars make their decision on whether they are going to apply for the funding, or how much time they are going to spend for the application which determines the qualities of their proposals. They receive a certain payoff based on whether their grant application will be funded or not. We will initially start from the simplest case focusing on an isolated scholar. We then gradually build up the model by adding other scholars and interaction between scholars into the model.

Isolated scholar

Suppose there is one scholar, named scholar i. Following Gross & Bergstrom (2019), we assume that scholar i has a certain quality q_i and can submit a proposal with strength s_i. The cost that it takes a scholar of quality q_i to submit a proposal with strength s_i is denoted by $c(q_i, s_i)$. The application of scholar i with a strength s_i is funded with probability $a(s_i)$. If the application is funded, he receives a payoff $y(q_i, s_i)$, the payoff depends on the proposal strength s_i since a good proposal may improve the actual research. If the application is not funded, he receives no payoff.

The payoff $y(q_i, s_i)$ can be interpreted as a combination of funds, prestige, career position, et cetera, that may result from receiving funding. The cost $c(q_i, s_i)$ can be interpreted as the time spent into preparing the application. If scholar i would not prepare a grant application he may do research instead, which could be interpreted as also providing a certain payoff to scholar i. Such opportunity costs can be considered to be part of the cost of preparing a grant application, and we assume this is included in the cost function $c(q_i, s_i)$.

The expected payoff of scholar i applying for funding is

$$a(s_i)y(q_i, s_i) - c(q_i, s_i).$$

To simplify the problem, we assume that the probability that the funding application is funded is dominated by a specific polynomial term,

$$a(s_i) = a_0 s_i^k,$$

where $a_0 > 0$ is some constant such that $a(s_i) \leq 1$ and k can be interpreted as the accuracy of the funding allocation process with respect to the proposal strength. The higher k, the larger the difference between the success probability of high-quality scholars and that of low-quality scholars.

Similarly, for the cost of writing the grant application, we may assume that

$$c(q_i, s_i) = c_0 s_i^{l_s} q_i^{l_q},$$

where c_0 is again a normalizing constant and l_s and l_q represent how costs scale with strength and quality.

Finally, we assume that the payoff from a successful application is simply proportional to the scholar's quality q_i

$$y(q_i, s_i) = y_0 q_i,$$

where y_0 can be interpreted as an abstract representation of the amount of funding.

Under these simplifying assumptions the expected payoff of scholar i is

$$E_{single} = a_0 s_i^k y_0 q_i - c_0 s_i^{l_s} q_i^{l_q}.$$

The astute reader may observe that a_0 and y_0 are degenerate in the model. We retain them here separately to clarify the role of each variable.

To maximize the expected payoff, scholar i would prepare a proposal with a strength s_i such that

$$\frac{d(a(s_i)y(q_i, s_i) - c(q_i, s_i))}{ds_i} = 0,$$

which implies

$$s_i = \left(\frac{c_0 l_s q_i^{l_q - 1}}{a_0 y_0 k}\right)^{\frac{1}{k - l_s}}.$$

The optimal proposal strength s_i is proportional to $q_i^{\frac{l_q - 1}{k - l_s}}$ Fig. 1 (left) visualizes the optimal proposal strength as a function of scholar's quality under different sets of parameters.

We are interested in studying the efficiency of the funding system. There are various possible definitions of efficiency. Gross and Bergstrom (2019) define efficiency as the ratio of the expected gain and the expected cost. Here we use an alternative definition of efficiency, since it is possible to have zero expected cost in some models and the efficiency is not well normalized when the cost expected cost is close to zero.

We propose the following definition of efficiency:

- Perfect allocation configuration: only scholars with the highest payoff apply for the funding, see Fig. 1 for an example.
- Efficiency: ratio of the expected payoff of the system and the payoff from the perfect allocation configuration.

Our definition of efficiency is meaningful only for models with more than one scholar, so we do not calculate it for the isolated scholar model. On the other hand, if we assume there is a distribution of the scholars, the efficiency can be computed, see (Lai, 2021) for the solution.

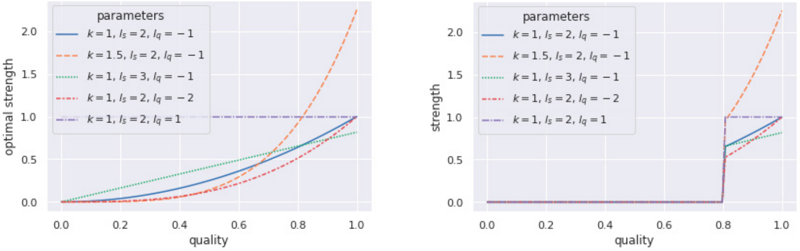


Figure 1. Isolated scholar model with parameters $a_0 = 1$, $y_0 = 2$, $c_0 = 1$, (left) optimal proposal strength versus scholar's quality, (right) the perfect allocation configuration where only some high-quality scholars submit their proposals

Two scholars, discrete choice

In reality, scholars do not apply for funding in isolation. Rather, multiple scholars typically have to decide simultaneously to apply for funding. If more scholars apply, this typically decreases the overall chances for an individual scholar to receive funding. We now add one additional scholar to the model. Although in reality there will be more than two scholars applying for funding, limiting it to two scholars still enables us to obtain some analytical understanding. At the same time, this simple model allows us to understand how interaction between two scholars affects the efficiency of the system.

The probability of receiving funding depends on the number of applicants and their proposal strength. In the model of an isolated scholar, we derived an expression for the optimal proposal strength. However, in the two-scholar model, there are two proposal strengths involved, leading to a non-linear differential equation system which is not easy to solve. For simplicity and analytical tractability, we therefore assume that scholars choose between two options: applying for funding or not. In case they do apply for funding, we assume the cost of applying is proportional to their quality, i.e., $s_i = q_i$. This is motivated by the fact that the optimal proposal strength increases with scholar's quality in the isolated scholar model (see Fig. 1). Though the relation is not exactly linear in the isolated scholar model, $s_i = q_i$ can be viewed as an approximation to the case where the scholars follow the suggestion from the isolated scholar model. In case they do not apply for funding, we may assume that no costs are incurred. This can be achieved by a proposal strength $s_i = 0$. In short, scholars will choose between $s_i = 0$ and $s_i = q_i$. The setting reduces the system into a simple game theoretical system.

Consider a model of 2 scholars, where for each scholar $i = 1{,}2$

$$a(s_i) = a_0(s_1, s_2)s_i^k,$$
$$y(q_i, s_i) = y_0 q_i,$$
$$c(q_i, s_i) = c_0 s_i^{l_s} q_i^{l_q}.$$

Suppose there is only one grant available, for which the two scholars compete. Then if at least one scholar applies the probability to be funded is $a(s_i) = \frac{s_i^k}{s_1^k + s_2^k}$. Clearly, if only one scholar applies, that scholar is guaranteed to be funded. The model can be expressed as a normal-form game, where scholar 1 is the row player, and scholar 2 is the column player.

The payoff matrices for scholar 1 and 2 are given by

$$\begin{array}{cc} & \begin{array}{cc} N & A \end{array} \\ \begin{array}{c} N \\ A \end{array} & \begin{pmatrix} 0 & 0 \\ y_0 q_1 - c_0 q_1^l & \frac{q_1^k}{q_1^k + q_2^k} y_0 q_1 - c_0 q_1^l \end{pmatrix} \end{array}, \quad \begin{array}{cc} & \begin{array}{cc} N & A \end{array} \\ \begin{array}{c} N \\ A \end{array} & \begin{pmatrix} 0 & y_0 q_2 - c_0 q_2^l \\ 0 & \frac{q_2^k}{q_1^k + q_2^k} y_0 q_2 - c_0 q_2^l \end{pmatrix} \end{array}.$$

where $l = l_s + l_q$. A positive l can be interpreted as scholars spend their time on something to receive a payoff that is proportional to their quality, while a negative l can be interpreted as higher quality scholars less time or less effort to prepare a grant proposal.

The first row corresponds to the choice of scholar 1 to *not* apply (N), while the second row corresponds to the choice of scholar 1 to apply (A). The first column corresponds to the choice of scholar 2 to *not* apply (N), while the second column corresponds to the choice of scholar 2 to *not* apply (A).

When a scholar applies, we say he follows strategy A, while if he does not apply, he follows strategy N.

Strategy N is a dominant strategy for scholar 1 if

$$y_0 q_1 \leq c_0 q_1^l,$$

while strategy A is a dominant strategy for scholar 1 if

$$\frac{q_1^k}{q_1^k + q_2^k} y_0 \geq c_0 q_1^{l-1}.$$

Similarly, strategy N is a dominant strategy for scholar 2 if

$$y_0 q_1 \leq c_0 q_1^l,$$

while strategy A is a dominant strategy for scholar 2 if

$$\frac{q_2^k}{q_1^k + q_2^k} y_0 \geq c_0 q_2^{l-1}.$$

If no strategy is dominant, the system is characterized by a mixed Nash equilibrium with 0 payoff.

At every Nash equilibrium, whether it is a pure or mixed equilibrium, we can compute a payoff $E_{2-scholar}$. If at least one scholar applies, we can define the expected payoff from the perfect funding allocation configuration as

$$E_{perfect2-scholar} = max\left(y_0 q_1 - c_0 q_1^l, y_0 q_2 - c_0 q_2^l\right),$$

where only the scholar with the highest payoff applies for the funding.

We then define the efficiency as

$$\epsilon_{2-scholar} = \frac{E_{2-scholar}}{E_{perfect2-scholar}}.$$

If the dominant strategy for both scholars is to not apply, we define the efficiency to be 0.

n scholars, discrete choice, unlimited number of grant per scholar

To extend the two-scholars model to a n scholar model, suppose there are n scholar and m grants available. We here allow scholar i to receive all m grants to produce a simple model to compare with the two-scholars model. Similar to the previous section, we assume that scholar i can only choose between $s_i = 0$ or $s_i = q_i$. Let Q_A be the set of all qualities of scholars who choose to apply, i.e., $Q_A = \{i \mid s_i = q_i\}$. We have

$$a(s_i) = 0 \text{ or } \frac{m s_i^k}{\sum_{j \in Q_A} \left(s_j^k\right)},$$
$$y(q_i, s_i) = y_0 q_i,$$
$$c(q_i, s_i) = 0 \text{ or } c_0 q_i^l,$$

note that $a(s_i)$ here can be interpreted as the expected number of grants received, rather than the probability of receiving a grant.

Unfortunately, analytical results are difficult to obtain for this model, and we therefore rely on the polynomial weight algorithm (Roughgarden, 2010) to try to find equilibria. Though learning algorithms may not be able to find the Nash equilibrium (Daskalakis et al., 2010), the equilibrium found by the algorithm may still be realistic if we relate the models with the real world.

After simulating the strategy choices and payoffs of individual scholars, the expected total payoff $E_{n-scholar,unlimitedgrant}$ can be obtained by adding payoffs from all scholars and then take an average over a few steps. As before, the perfect allocation configuration is that only scholars with the highest quality q_h apply for the funding while others do not. The perfect allocation payoff is then

$$E_{perfectn-scholar,unlimitedgrant} = m y_0 q_h - c_0 q_h^l.$$

We again define efficiency as the ratio between the actual expected total payoff and the perfect allocation payoff, which we take to be zero if no scholars apply at all.

n scholars, discrete choice, one grant per scholar

In the previous model, we allow scholars to accumulate an unlimited number of grants. Although this model has a relatively straightforward definition of the probability to be funded (which is actually an expectation), most funding schemes only allocate a single grant to a scholar. We assume there are m grants available, where $m \leq n$, and each scholar can only get one grant at most. We again assume a discrete choice, and scholar i can choose $s_i = 0$ or $s_i = q_i$. We then obtain

$$\begin{aligned} a(s_i) &= 0 \text{ or } A(q_j \in Q_A, k), \\ y(q_i, s_i) &= y_0 q_i, \\ c(q_i, s_i) &= 0 \text{ or } c_0 q_i^l, \end{aligned}$$

where Q_A is the set of all qualities of scholars who choose to apply as before.

We again need to solve the mixed strategies of this model computationally. Since the success probability $A\left(q_j \in Q_a, k\right)$ of the one grant per scholar model is related to the weighted random sampling with a reservoir problem (Efraimidis & Spirakis, 2006), it is not efficient to compute the probability and the payoffs. Therefore, we use the partial information model (Roughgarden, 2010) to simulate the efficiency in a more agent-based modeling flavour.

The expected total payoff $E_{n-scholar,1grant}$ is computed by averaging the sum of individual payoffs for a few time steps. Let Q_h be the set of m highest scholars' qualities. The perfect allocation payoff is then

$$E_{\text{perfect n-scholar, 1 grant}} = \sum_{q_j \in Q_h} \left(y_0 q_j - c_0 q_j^l\right).$$

We again define efficiency as the ratio of the two.

Results

We have constructed three models:

- a two-scholar model,
- a n-scholar unlimited grant, and
- a n-scholar one grant per scholar model.

We now analyse how the efficiency in these three different models depend on the total amount of funding. The total amount of funding is represented by the variable my_0 ($m = 1$ for the two-scholar model), and the efficiency is defined as the ratio of the expected total payoff and the payoff from the perfect allocation configuration. For each model, we present results for some specific sets of parameters. If we have any analytical results, we will simply present those. Alternatively, if we are unable to obtain analytical results, we will present numerical approximations from simulations. In that case, we always perform 10 runs of the simulations and present averages and deviations thereof.

Efficiency in the two-scholars model

There are 6 free variables in the two-scholars model. Without loss of generality, we assume that $q_2 > q_1$. We choose $q_1 = 0.3$, $q_2 = 0.7$, and $c_0 = 1$. We set $k = 1$, so that the probability of being funded is proportional to a scholar's quality. We set $l = 1$, so that the cost of applying is proportional to a scholar's quality.

Since the payoffs received by scholar 1 and 2, if both of them apply for the funding, are closely related to their dominant strategies A, we plot Fig. 2 to show how those payoffs vary with the amount of funding y_0, if the payoff is greater than 0, then always applying is a dominant strategy for the scholar.

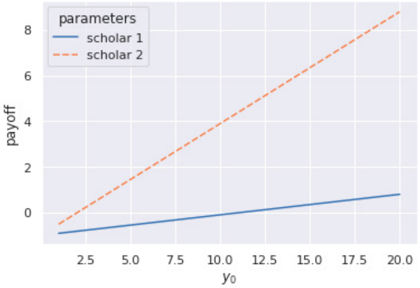


Figure 2. The payoffs scholar 1 and 2 receive as a function of the amount of funding y_0, assuming that both scholar 1 and 2 apply for the funding.

Multiple Nash equilibria may exist in a 2-player game, where they may have different efficiencies. We use the two equilibria with the highest and lowest efficiency to bound the efficiency in this model. Fig. 3 plots the efficiency of those equilibria as a function of the amount of funding y_0. It is interesting to note that we can discern three distinct phases, a feature that does not exist in the single scholar model with a distribution (Lai, 2021). The first phase corresponds to a "no dominant strategy" region, where mixed Nash equilibrium with 0 payoff is allowed. Increasing the funding y_0 we transition towards the second phase, where the dominant strategy for scholar 2 is to always apply for funding, which forces scholar 1 to not apply for funding (because $q_2 > q_1$). This produces a 100% efficient configuration. As the amount of funding further increases, we transition towards the third phase, where applying for funding also becomes a dominant strategy for scholar 2, leading to a sudden drop in efficiency.

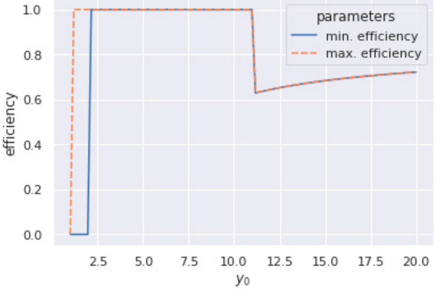


Figure 3. Maximum and minimum efficiency in the two-scholars model as a function of the amount of funding y_0.

The middle phase, which is most efficient, can be enlarged by increasing k or by decreasing l. Increasing k means that the higher quality scholar is more likely to receive funding if both scholars apply. We can interpret this as the accuracy of peer review, in the sense that peer review would be expected to be able to identify the higher quality scholar. In that case, the lower quality scholar requires an even larger funding amount y_0 in order to also apply (thereby entering the third phase), because the probability he is funded is lower with higher k. Increasing k therefore shifts the threshold between the second and third phase to higher funding amounts y_0. On the other hand, the higher quality scholar has a higher probability to get funded, so a smaller y_0 is needed to reach the dominant strategy A, thus the threshold between the first and second phase shifts to lower funding amount y_0.

Decreasing l means that the cost of applying becomes relatively lower for higher quality. In this case, the lower quality scholar requires an even larger funding amount y_0 in order to also apply (thereby entering the third phase), because the relative cost of applying is higher with lower l. Increasing l therefore shifts the threshold between the second and third phase to higher funding amounts y_0.

The two-scholar model shows that increasing the funding beyond a certain threshold has a counter-intuitive effect of decreasing the funding efficiency. This is because increasing funding leads the lower quality scholar to also apply, which incurs a cost, but does not entail additional benefits.

Efficiency in the n scholar unlimited grant model

Without considering the distribution of scholars' quality, there are 6 free parameters. Following the previous sections, we set $k = 1$, $c_0 = 1$, and $l = 1$.

We use a non-random, fixed-interval uniform distribution $q_{min} \le q_i \le q_{max}$ for some exploration work, the quality of scholar i is always $q_{min} + \frac{i-1}{q_{max}-q_{min}}$. Subsequently, we use a beta distribution to be more flexible while being compatible with the analytical solution in (Lai, 2021). We use the notation $q_i \sim \mathrm{Beta}(\alpha, \beta)$ to indicate that q_i follows a beta distribution.

We first attempt to reproduce the three phases that we observed in the two-scholar model. We set the number of scholars in the model to $n = 2,6,20,50$ and the number of grants to $m = 1$. We also assume that the scholars' qualities follow a fixed-interval uniform distribution with a minimum of 0.3 and maximum of 0.7, making the $n = 2$ case identical to the two-scholar model. We simulate the model and compute the efficiency. Fig. 4 shows that this model indeed reproduces a similar pattern as the two-scholar model. While the patterns are similar, we can see that the exact Nash equilibria are not reached in the simulations so that the patterns are not exactly the same. The efficiency decreases when we increase the scholars since the competition is higher and the total available resource remains unchanged.

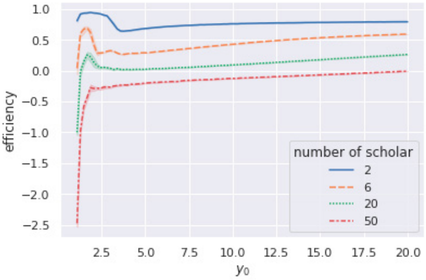


Figure 4. The efficiency of the n scholar unlimited grant model as a function of the amount of funding y_0, with $n = 2,6,20,50$, scholars' qualities follow uniform distribution.

We now explore the model further with $n = 100$ scholars, with beta distributed qualities. Fig. 5 shows the plot of the efficiency as a function of y_0 for $q_i \sim \mathrm{Beta}(2,2)$. The results show that having a higher k (more accurate assessment of scholars' quality) is clearly beneficial, while having a negative l does not necessarily improve the efficiency. The results also depend on the distribution of scholars' qualities. Fig. 6 shows results with $q_i \sim Beta(10,2)$, which means the qualities of scholars are in general higher. Interestingly, the efficiency is not higher than when using $q_i \sim Beta(2{,}2)$,g . This suggests that having more high-quality scholars does not necessarily imply a higher efficiency since the competition is also tougher.

In every simulation, there is a steep increase in efficiency when y_0 is small. It is because better scholars are easier to have higher payoffs when choosing to apply and they are more likely to apply, thus discouraging lower-quality scholars to apply. For some parameter settings, there is a peak in the efficiency curve, similar to the three phases we observed in the two-scholars model, but this does not happen for all parameter settings.

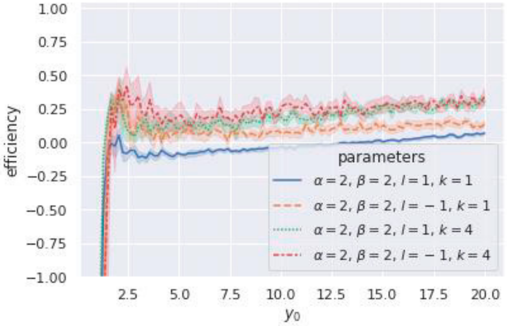


Figure 5. The efficiency of the $n = 100$ scholar unlimited grant model as a function of the amount of funding y_0, scholars' qualities follow the beta distribution with $\alpha = 2$ and $\beta = 2$.

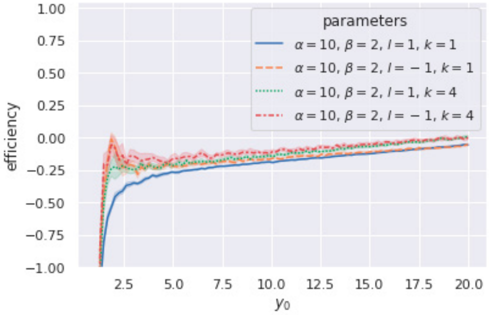


Figure 6. The efficiency of the $n = 100$ scholar unlimited grant model as a function of the amount of funding y_0, scholars' qualities follow the beta distribution with $\alpha = 10$ and $\beta = 2$.

Efficiency in the n *scholar one grant per scholar model*

Similar to the unlimited grant model, excluding the parameters about the quality distribution, there are 6 free parameters. We take $k = 1$, $c_0 = 1$, $l = 1$ as the starting case. Fig. 7 shows the efficiency curve with $n = 2{,}6{,}20{,}50$ and $m = 1$. We can observe the 3-phase feature for $n = 2$, while for larger n ($n = 20$, $n = 50$) the peak disappears. While it is possible that this is the nature of the model, insufficient steps in the simulations may also cause this observation since the algorithm here needs more time steps to converge to equilibria comparing to the algorithm used in the unlimited grant model. Due to the limitation of computational power, we leave this for future studies.

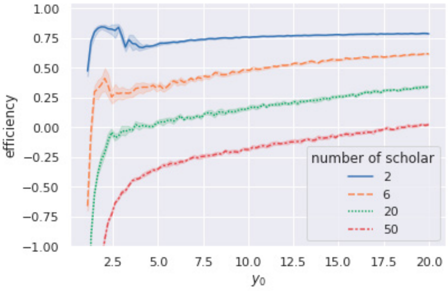


Figure 7. The efficiency of the $n = 2,6,20,50$ scholar one grant per scholar model as a function of the amount of funding y_0, scholars' qualities follow uniform distribution.

To investigate how different parameter settings in the model affect the efficiency, we run simulations with $n = 100$ scholars and beta distributed scholars' qualities. Unlike the unlimited grant model, the number of funding m is a meaningful variable in this model which is not degenerate with y_0. However, in our exploratory simulations, we don't observe any effects of increasing m besides having a higher overall efficiency, which is similar to the effect of increasing y_0. We therefore fix $m = 20$ such that at most 20 scholars receive funding. Fig. 8 shows the efficiency curves with $l = 1 \ or - 1, k = 1 \ or \ 4, \alpha = 2 \ and \ \beta = 2$. Fig. 9 shows a similar result but with $\alpha = 10$. Despite the rapid increase in the efficiency when y_0 is small, there is no 3-phase we observed in some of the earlier plots. We again observe that increasing k seems to increase the efficiency while the effect of varying l is less clear. In contrast to the result in the unlimited grant model, where $\alpha = 2$ simulations are having a much higher efficiencies than that in the unlimited grant model, the efficiencies for $\alpha = 2$ and $\alpha = 10$ are closer in this model.

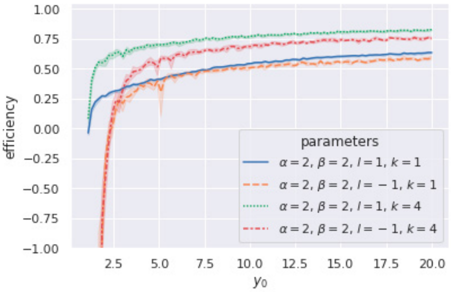


Figure 8. The efficiency of the $n = 100$ scholar one grant per scholar model as a function of the amount of funding y_0, scholars' qualities follow the beta distribution with $\alpha = 2$ and $\beta = 2$.

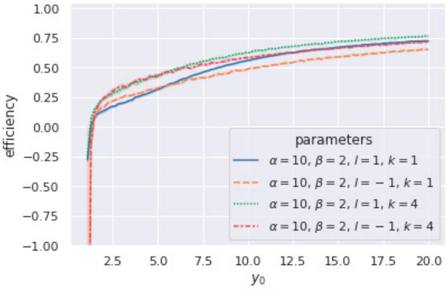


Figure 9. The efficiency of the $n = 100$ scholar one grant per scholar model as a function of the amount of funding y_0, scholars' qualities follow the beta distribution with $\alpha = 10$ and $\beta = 2$.

Discussion

We presented three different simple model of funding in science. Although are models are highly stylized, the results may still provide some insight. We here summarize our understanding of the results, and we discuss future research questions and improvements to the models.

Regardless of the model, if the amount of funding is very large, the efficiency almost always approaches 1. This is easy to see because the costs remain fixed, becoming negligible in the limit of large amounts of funding. On the other hand, in many of the scenarios, we are able to identify three distinct phases, that showed distinct dynamics. The first phase is characterised by both low and high quality scholars not having a dominant strategy. When increasing the funding, this transition towards the second phase, where only high quality scholars apply for a grant, leading to a highly efficient funding system. Increasing the funding further, transitions the system towards a third phase, where lower quality scholars again also start to apply for a grant, thereby decreasing the efficiency of the funding system. In scenarios where we cannot discern these three phases, we still observe a steep initial rise in efficiency when increasing funding, which levels off for further increases in funding. This suggest that increasing the amount of funding in the system may yield only modest increases in efficiency.

In all settings considered in this study, having a larger accuracy of peer review seems to help to increase efficiency. That is, if peer review is better able to identify high quality scholars, this increases efficiency. In contrast, how the cost of applying scales with a scholar's quality, seems to have less clear consequences for efficiency. Our results suggest that it would be beneficial to improve the review process of funding application, while (re)-designing the application procedure to modify the cost is less relevant. Of course, the accuracy of peer review may be something that is inherently limited.

In our results, we find that some high-quality scholars are needed in order to achieve an efficient funding system. However, too many high-quality scholars may create unnecessary competition and may reduce the efficiency. Moreover, the quality of a scholar cannot simply be modeled as a single dimensional variable. Quality may multi-faceted with dimensions such as creativity, novelty, or rigour. A more realistic model may also include such alternative dimensions of quality.

The funding allocation itself also affects efficiency. When allowing researchers to accumulate unlimited grants, we see that the efficiency is substantially lower than in grant models in which each scholar can receive only a single grant. In reality, the number of grant that a scholar can apply is limited. Having a clever design of the number and the size of grants available to different scholars can help improving the efficiency.

There are several limitations to our models. Firstly, the cost can depend on the amount of funding. It is natural that scholars spend more effort on preparing for a larger grant than for a smaller one. Although we capture this behaviour in the isolated scholar model, we ignored it later to simplify the models. Secondly, the payoff of receiving funding does not necessarily translate immediately to the amount of funding that is received. The interpretation of y_0 as the funding amount can therefore be discussed. Thirdly, it may be argued that the total amount of funding is simply a given, and that it cannot be changed. However, it is possible that the amount of funding for one funding scheme is increased, while other types of funding are decreased. Despite some unrealistic factors, we think our observations highlight what might be possible when we change the amount of available funding.

References

Adam, D. (2019). Science funders gamble on grant lotteries. *Nature,* 575(7785), 574–5.

Avin, S. (2015). Funding science by lottery. Recent developments in the philosophy of science: Epsa13 Helsinki (pp. 111–126). Springer.

Carson, L., Bartneck, C. & Voges, K. (2013). Over-competitiveness in academia: A literature review. *Disruptive Science and Technology*, 1(4), 183–190.

Daskalakis, C., Frongillo, R., Papadimitriou, C. H., Pierrakos, G. & Valiant, G. (2010). On learning algorithms for Nash equilibria. *International Symposium on Algorithmic Game Theory,* 114–125.

Efraimidis, P. S. & Spirakis, P. G. (2006). Weighted random sampling with a reservoir. *Information Processing Letters*, 97 (5), 181–185.

Fang, F. C. & Casadevall, A. (2016). *mBio* 7(2), e00422-16.

Geard, N. & Noble, J. (2010). Modelling academic research funding as a resource allocation problem. *3rd World Congress on Social Simulation.*

Gross, K. & Bergstrom, C. T. (2019). Contest models highlight inherent inefficiencies of scientific funding competitions. *PLoS Biology*, 17(1), e3000065.

Lai, K. H. (2021). On the analytic solution of a no-interaction research system mode https://zenodo.org/record/4540375#.YIrPTbUzZMIl.

Liu, M., Choy, V., Clarke, P., Barnett, A., Blakely, T. & Pomeroy, L. (2020). The acceptability of using a lottery to allocate research funding: A survey of applicants. *Research Integrity and Peer Review*, 5(1), 3.

Roberts, P. (2007). Neoliberalism, performativity and research. *International Review of Education*, 53(4), 349–365.

Roughgarden, T. (2010). Algorithmic game theory. *Communications of the ACM*, 53(7), 78–86.

An appraisal of publication embedding techniques in the context of conventional bibliometric relatedness measures

Wout S. Lamers[1,*], Nees Jan van Eck[1] and Giovanni Colavizza[2]

[1] *{w.s.lamers, ecknjpvan}@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies (CWTS), Leiden (The Netherlands)

[2] *g.colavizza@uva.nl*
University of Amsterdam, Faculty of Humanities, Department Mediastudies, Amsterdam (The Netherlands)

Abstract

Modern natural language processing techniques have given rise to embedding techniques that can represent documents based on their content or context, and several papers have operationalized these to perform bibliometric tasks. The relationship between these embeddings and conventional citation based or title and abstract based mappings remains unclear. Contrary to citation-based or term-based relatedness, embedding-based relatedness is not immediately interpretable. We consider four embedding-derived publication relatedness measures, based on: 1) word2vec embeddings of citation labels, sentence embeddings using 2) BERT and 3) SciBERT, and 4) title and abstract embeddings using SPECTER, and compare them with conventional bibliometric publication relatedness measures derived from citation relations and title and abstract noun phrases. We show that there is stronger overlap between these embedding-derived relatedness measures and citation-based relatedness than with title and abstract noun phrase-based relatedness, and that embedding-derived relatedness measures outperform conventional techniques when used to cluster publications cited with the same citation intent.

Introduction

The notion of relatedness of scientific publications is at the heart of many open problems in bibliometrics. Publication clustering and mapping techniques rely on publication-to-publication relatedness measures, conventionally based on citation relations, textual similarities of titles and abstracts, or occasionally combinations of these. It should come as no surprise that different relatedness measures produce different results (Gläser et al., 2017), which makes the informed choice of appropriate relatedness measures an imperative aspect of many bibliometric research tasks.

Meanwhile, advances in natural language processing (NLP) and information retrieval have introduced a host of new text-based methods for establishing relatedness of not only documents to one another, but also documents to terms in the vocabulary of these documents. The introduction of word2vec (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) marked a substantial advance, allowing vector representations of words to be learned that not only allow for easy computation of semantic similarities, but also preserve—to some extent—the semantic and syntactic relationships between terms. Doc2vec (Le & Mikolov, 2014) adapted this method, allowing for the generation of similar vector representations for larger bodies of text. Further advances in the field since then such as the emergence of large pre-trained language models like BERT (Devlin et al., 2018) have only expanded the suite of options for establishing relatedness between documents.

The case for using NLP techniques in science mapping and clustering is compelling. At their core lies the distributional hypothesis, the assertion that terms with similar distributions have similar meaning, i.e., similar terms are used in similar context. What is citation, if not the deliberate act of putting previous work in the context of a new author's design? Hence, papers that are cited in similar context, or that introduce similar language, can be assumed to be related to each other, and embedding techniques can capture this relatedness. Another important benefit of these embedding methods is that publications can be embedded along with terms, allowing researchers to learn relationships between embedded literature and concepts in the vocabulary, directly from the data.

Several recent publications have leveraged these new NLP techniques to explore new ways of representing publications, visualize research landscapes, and facilitate document search and retrieval. Approaches range from representing documents using the text in their abstracts (e.g. Hain et al., 2020) to embeddings derived from citation context (e.g. Berger et al., 2017; He & Chen, 2018), and a number of studies have explored combining text-based approaches with citation-based approaches (e.g. Cohan et al., 2020; Ganguly & Pudi, 2017). However, while many of these papers find that their representations of scientific literature capture useful and relevant aspects of publication context, topic, purpose, or overall relatedness, and often outperform traditional techniques in specific tests selected by the authors, embedding-derived relatedness measures remain difficult to interpret, and their resulting vector representations of papers are essentially black boxes. The nature of the resulting relatedness measures is difficult to compare to traditional bibliometric techniques, and this ambiguity makes the use of these new embedding techniques in science mapping and publication clustering less attractive.

Our research aims to compare a set of common NLP-based publication representations and their resulting relatedness measures with traditional citation-based and abstract-based publication relatedness measures. We follow the methodology introduced by Waltman et al. (2020), allowing us to compare the accuracy of clustering results obtained using various publication relatedness measures based on a selected baseline relatedness measure. We compare four methods for generating publication embeddings with two established bibliometric relatedness measures, one based on publication titles and abstracts, and one based on citations, in two data sets. Our goal is to establish whether embedding-derived publication relatedness measures more strongly resemble established text-based relatedness measures or citation-based relatedness measures.

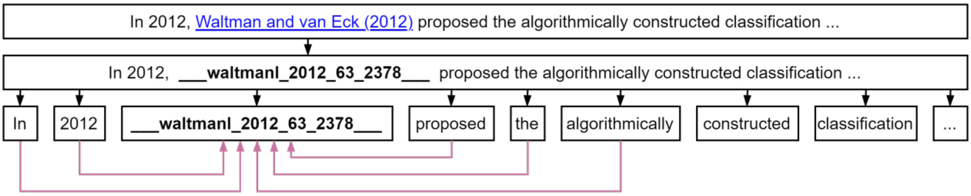
Methods

Our first data set is selected from the Elsevier ScienceDirect corpus previously analyzed in detail by Boyack et al. (2018). We select scientometrics-related publications using the meso-level of the CWTS hierarchical publication classification system (Waltman & van Eck, 2012). This produced a total of 7825 publications, that are cited in 51003 full-text sentences (not limited to the scientometrics field). 3169 of these publications feature in at least five citing sentences, for a total of 44464 citing sentences. If a sentence features multiple citations, only the first citation is considered, to avoid associating the exact same context with multiple works.

The second dataset is the SciCite citation intent dataset introduced by Cohan et al. (2019). This consists of 6427 publications and 10817 sentences citing these publications, as well as annotated citation intent labels noting whether a citing sentence cites a paper as *background*, for its *methods*, or to compare *results*. Titles and abstracts of these publications, as well as publications cited by or citing them, were retrieved using the Semantic Scholar API.

Word2vec

Our first embedding-derived relatedness measure is a simple word2vec embedding of citation labels. In each citing sentence, the label referencing the cited publication is replaced with a token unique to the cited work, and a word2vec embedding is trained on the resulting corpus using the *gensim* python library (Rehurek & Sojka, 2010). For a schematic representation, see figure 1. Relatedness of publications is calculated as the cosine similarities of the unique publication tokens. The word2vec architecture has a few adjustable hyperparameters. We varied the embedding size, window size, epochs, and negative sample size, and tried both the CBOW and skip-gram methods. Skip-gram with a window size encompassing the entire sentence, keeping other parameters set to their default values, produced the most stable publication-to-publication similarities over multiple model runs.



In 2012, Waltman and van Eck (2012) proposed the algorithmically constructed classification ...
In 2012, ___waltmanl_2012_63_2378___ proposed the algorithmically constructed classification ...
In
2012
___waltmanl_2012_63_2378___
proposed
the
algorithmically
constructed
classification
...

Figure 1: example of word2vec embedding of citation labels, window size three.

Sentence-BERT

Entire sentences can be embedded using Sentence-BERT (Reimers & Gurevych, 2019). Averaging the embeddings of sentences that cite the same paper then offers us another way to generate representations for cited publications based on their citation context. In this procedure, we again associated sentences only with the first cited publication, and omitted citation labels entirely from the sentence. Since the Sentence-BERT architecture can use a variety of language models, we repeated this process twice, once using the original BERT base model (Devlin et al., 2018), and once using the SciBERT model (Beltagy et al., 2019) which was trained specifically on scientific text.

SPECTER

Our final embedding-based relatedness measure uses SPECTER (Cohan et al., 2020) to generate publication representations based on titles and abstracts alone. SPECTER's core feature is that it generates document-level representations using only titles and abstracts, in a manner informed by pre-training on a citation graph. This allows it to outperform other architectures that rely solely on title and abstract information in a variety of performance benchmarks such as citation prediction and topic classification (Cohan et al., 2020).

Conventional relatedness measures and comparison

Waltman et al. (2020) introduce a methodology for comparing publication relatedness measures based on the accuracy of clustering solutions, compared to a baseline relatedness measure. Following their work, we use the BM25 text similarity measure based on title and abstract noun phrases, and the combined direct citation, bibliographic coupling and co-citation (DC-BC-CC) relatedness measure, to represent baseline text-based and citation-based relatedness measures, respectively. All relatedness measures are also simplified and normalized, fist by keeping only the 20 most strongly related connections for each document, and subsequently normalizing their weights by dividing the relatedness of each outgoing edge by the total weight of outgoing edges for each node.

Waltman et al. (2020) then define the accuracy of a clustering solution based on relatedness measure X, evaluated against measure C, as

$$A^{X|C} = \frac{1}{N}\sum\nolimits_{ij}(c_i^X = c_j^X)r_{ij}^C$$

where $\left(c_i^X = c_j^X\right)$ equals 1 if the clustering c of document i is the same as that of document j, and 0 otherwise, with r_{ij}^C the relatedness between documents i and j per measure C. The normalization of the relatedness measures constrains this accuracy score between 0 and 1. Accuracy is computed for a variety of clustering solutions obtained using the Leiden algorithm (Traag et al., 2019), using a range of clustering resolution parameters, and the obtained accuracy is then plotted against the granularity of the resulting clustering solution in a granularity-accuracy plot, allowing us to easily compare the clustering accuracy of relatedness measures over a range of resolutions.

Results

Figure 2a displays the accuracy of the clustering solutions resulting from the various relatedness measures, using the BM25 title and abstract based relatedness as a baseline, over different clustering granularities. This analysis is limited to only the 2979 publications that are connected by relatedness in each of the six networks. Figure 2b uses the DC-BC-CC citation relatedness baseline, while 2c and 2d use word2vec and SPECTER-derived relatedness instead.

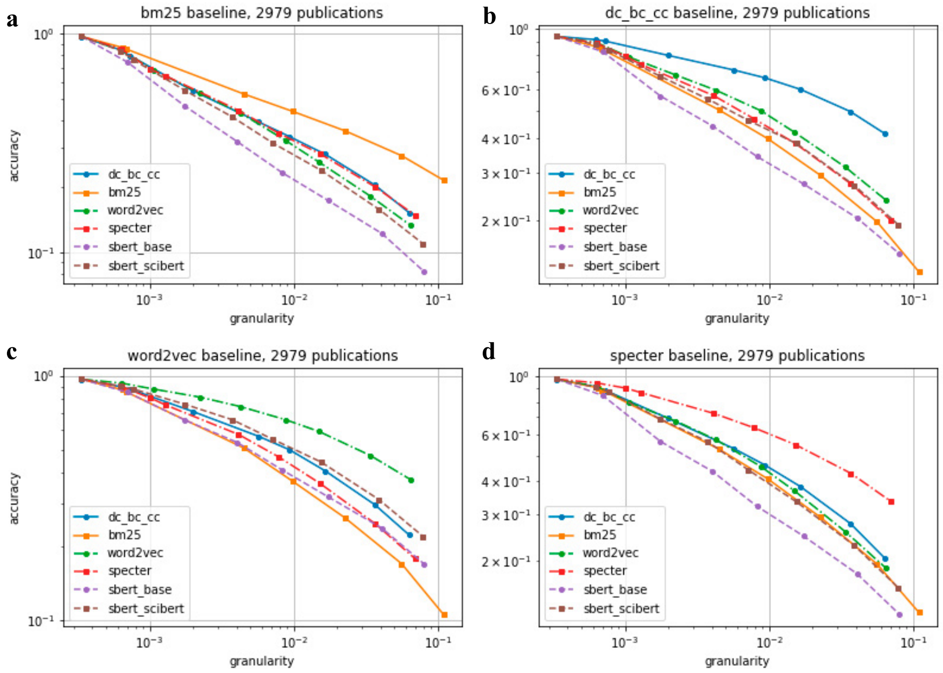


Figure 2: granularity-accuracy plots, scientometrics data set.

Figure 2c shows us that, in our scientometrics data set, the word2vec embedding derived from citing context resembles the citation-based relatedness measures more strongly than it does the title and abstract text-based relatedness measure. This is perhaps surprising, as it is fundamentally text-based, though this text originates from sentences that cite, which would explain the similarities to citation relatedness patterns. More surprising, figure 2d shows that the SPECTER-embeddings, despite being based on the same title and abstract text as the BM25 relatedness, behave more like the DC-BC-CC citation relatedness. Indeed, figure 2a shows that SPECTER and DC-BC-CC are approximately equally accurate when compared directly against the BM25 relatedness. Finally, figure 2b shows that word2vec, SPECTER and the SciBERT-based sentence embedding more closely resemble citation relatedness than BM25, which was previously found to be the best-performing title and abstract relatedness measure, at least in our sample in which publications with 5 or more in-text citations are considered. When also including publications with fewer than 5 in-text citations (not pictured), SPECTER and BM25 are the most similar to DC-BC-CC, followed by word2vec and SciBERT sentence embedding.

The SciCite dataset, moreover, contains not only citation context for each publication, but is hand-labeled with citation intent. These labels (*background*, *methods*, and *results*) make for a poor relatedness measure but can be assessed at the cluster level. We can express the diversity of these labels over the clusters using Rao-Stirling diversity (see e.g. Leydesdorff et al., 2019)

$$\Delta = \sum_{i,j} d_{ij}(p_i p_j)$$

in which d_{ij} is a distance measure between clusters i and j, and p_i is the proportion of nodes in cluster i. To compute a distance measure, we take the numbers of citing sentences labeled *background*, *methods* and *results* in each cluster as a three-length vector and take the cosine distance of these vectors between each cluster pair. This means that diversity is maximized if publications cited with the same intent are clustered together. We vary the clustering resolution and plot the resulting diversity of SciCite intent labels against cluster granularity in Figure 3.

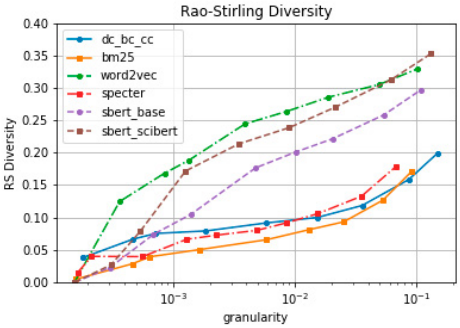


Figure 3: diversity-granularity plot of citation intent labels, SciCite dataset.

Figure 3 shows us that, with some margin, the citation context derived embeddings outperform the conventional relatedness measures, and that SPECTER improves only marginally over BM25 when it comes to clustering citation intent. Even BERT base, while lagging in figure 2, improves markedly on the conventional relatedness measures.

Discussion

Our research shows that embedding representations of publications have the potential to resemble conventional citation-based relatedness structures more accurately than the best-available conventional title and abstract-based relatedness measure. This holds true even for a simple word2vec embedding, provided enough citation context is available. Context-based embeddings also show a stronger correlation to citation intent labels in the SciCite data set than conventional citation-based or abstract-based relatedness measures.

It must be noted that the embedding-based relatedness measures used in this research might be further explored and optimized for the task of establishing publication relatedness. Especially BERT-based representations are typically fine-tuned for downstream tasks. Nevertheless, our research demonstrates the base feasibility of using embedding representations of publications and provides important insight into how they relate to traditional publication relatedness measures.

References

Beltagy, I., Lo, K. & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, 3615–3620.

Berger, M., Mcdonough, K. & Seversky, L. M. (2017). cite2vec: Citation-Driven Document Exploration via Word Embeddings. *IEEE Transactions on Visualization and Computer Graphics,* 23(1), 691-700.

Boyack, K. W., van Eck, N. J., Colavizza, G. & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, 12(1), 59–73.

Cohan, A., Ammar, W., Van, M. & Cady, Z. F. (2019). Structural Scaffolds for Citation Intent Classification in Scientific Publications. *Proceedings of NAACL-HLT*, 3586–3596.

Cohan, A., Feldman, S., Beltagy, I., Downey, D. & Weld, D. (2020). SPECTER: Document-level Representation Learning using Citation-informed Transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2270–2282.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Ganguly, S. & Pudi, V. (2017). Paper2vec: Combining graph and text information for scientific paper representation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10193 LNCS*, 383–395.

Gläser, J., Glänzel, W. & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, 111(2), 981–998.

Hain, D., Jurowetzki, R., Buchmann, T. & Wolf, P. (2020). *Text-based Technological Signatures and Similarities: How to create them and what to do with them*. http://arxiv.org/abs/2003.12303

He, J. & Chen, C. (2018). Temporal Representations of Citations for Understanding the Changing Roles of Scientific Publications. *Frontiers in Research Metrics and Analytics*, 3, 27.

Le, Q. V. & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *31st International Conference on Machine Learning, ICML 2014*, 4, 2931–2939.

Leydesdorff, L., Wagner, C. S. & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics*, 13(1), 255–269.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013, January 16). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26.

Rehurek, R. & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing,* 3982–3992.

Traag, V. A., Waltman, L. & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), 1–12.

Waltman, L., Boyack, K. W., Colavizza, G. & Van Eck, N. J. (2020). A principled methodology for comparing relatedness measures for clustering publications. *Quantitative Science Studies*, 1(2), 691–702.

Waltman, L. & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378–2392.

Exploring the Relationship between Qualities of Press Releases to Research Articles and the Articles' Impact

Steffen Lemke[1], Julian Sakmann[2], Max Brede[3] and Isabella Peters[4]

[1] *s.lemke@zbw.eu,* [2] *stu205601@mail.uni-kiel.de,* [3] *mbre@informatik.uni-kiel.de,* [4] *i.peters@zbw.eu*
[1, 2, 4] ZBW – Leibniz Information Centre for Economics, Düsternbrooker Weg 120, 24105 Kiel (Germany)
[3, 4] Kiel University, Christian-Albrechts-Platz 4, 24118 Kiel (Germany)

Abstract

Several studies have found that mentions of research articles in public media can have substantial effects on the articles' later citation counts and altmetrics. However, little attention so far went into investigating the potential relationship between qualitative properties of press texts that promote research and the research's impact. In this research in progress, we set out to manually analyze and compare the press releases published on *EurekAlert!* to promote a sample of 120 research articles, 60 of which later performed remarkably well concerning selected article-level metrics, while the remaining 60 articles later performed comparatively poorly. As a preliminary result, qualitative differences could be found regarding the press releases' structure, linguistic accessibility and the existence of narratives. First applications of our in-development codebook suggest associations between press releases with poor structure or accessibility and promoted research articles' metrics performance. We conclude with indications towards numerous promising paths for continuations of this study.

Introduction: Motivation and Related Work

In surveys on researchers' perceptions of bibliometric and altmetric indicators, a frequently encountered suspicion is that they might primarily capture visibility or curiosity, and therefore often ultimately the amount of effort made to advertise respective publications (Lemke et al., 2019; Nicholas et al., 2020). Besides efforts of *internal* science communication (i.e., communication primarily targeting other researchers, like for instance a presentation at an academic conference), this also includes the promotion research receives in channels of *external* science communication, e.g., by being featured in newspapers, podcasts, or television. Various studies analyzed the relationship between research publications' media visibility around the time of their publication and their later metrics, most often focusing on citation counts. For instance, several studies found newspaper coverage to be associated with substantially higher later citation rates for featured research articles (Phillips et al., 1991; Kiernan, 2003; Fanelli, 2013). Similarly, Chapman, Nguyen & White (2007) examined the association between articles published in the journal *Tobacco Control* receiving promotion in press releases and their later citations and usage metrics, finding the articles with accompanying press releases to be more likely to get cited, as well as to receive more downloads and web hits than similar articles without press releases. Lemke (2020) compared the citations and five prevalent altmetrics of a treatment group of 10,483 journal articles that were featured in press releases in 2016 to those of a similarly structured control group without known press release promotion, finding the treatment group to perform substantially better regarding all six examined indicators.

So while several previous studies revealed correlations between presence in different formats of external science communication and respective research articles' later metrics, it remains uncertain which processes and causalities explain these findings. Phillips et al. (1991) propose two hypotheses, the 'publicity hypothesis' and the 'earmark hypothesis'. The publicity hypothesis argues that it is the increase of visibility achieved by press release- or newspaper coverage that leads to more potential citers reading the featured articles and therefore increases their likelihood of receiving citations. The earmark hypothesis on the other hand suggests that the journalists selecting publications to cover and the researchers selecting publications to cite just independently of each other arrive at similar judgments regarding which literature suits their needs best. The citation advantage of a publication featured in mainstream media would therefore be the result of its own quality and not depend on the increased visibility. The results by Phillips et al. (1991) themselves provide a strong argument for the publicity hypothesis, as they found research articles featured in issues of the *New York Times* that had not been distributed due to a strike to receive considerably less citations than research articles featured in regular issues of the *New York Times*. However, the effects proposed by publicity hypothesis and earmark hypothesis are not necessarily mutually exclusive.

Despite the substantial number of studies evidencing an association between research articles' coverage in external science communication and later metrics, the 'hows' and 'whys' behind this association remain mostly unanswered. One reason for this could be most past studies' focus on purely quantitative relationships between mentions in external science communication and later citations, which typically regarded an article receiving press promotion as a numerical value or even as a binary event (i.e., without differentiation between promoting material's qualities). Potential structural or content-related properties of the actual press texts that promoted different articles on the other hand were mostly neglected so far.

This research in progress aims to address this gap through the qualitative analysis of press releases promoting research articles. By retrospectively analyzing press releases issued for journal articles that later received particularly high metrics and comparing these to press releases for journal articles published around the same time that later received comparatively low metrics, we aim to examine whether qualitative differences between the two groups of press releases are distinguishable. Our endeavor is guided by the hypothesis that PR activities, such as press releases, and press coverage can substantially affect research's overall impact, and that examining qualitative properties of said PR's and press coverage's individual instances could lead to a better understanding of the circumstances under which this is the case. As this article describes a study that is still in progress, its main purposes are twofold: (1) it intends to shed light on promising avenues for further research on the subject (and more specifically outline subsequent steps our own research will take), and (2) it shall give first insights into noteworthy observations we made during the manual coding of press releases to scholarly articles so far.

Within the framework of the news value theory, a substantial body of research already discussed the question which factors increase the likelihood for a topic to receive coverage in public media (see Badenschier and Wormer (2012) for a brief review of such works, as well as for a model explaining news factors for the particular case of science topics). Such factors (e.g., a topic's range, actuality, or surprise factor) explain why certain topics and therefore research articles might be selected for press releases in the first place. To complement this existing research, our main focus is on further properties of the press releases that should be largely independent of the promoted research's topic. In particular, we investigate whether particularly 'impactful' articles' press releases vary from others regarding their structure, linguistic accessibility, and the way they use emotionally engaging narratives to report the featured research's findings. Thus, our main interest is to describe how external science communicators, e.g., press officers of research institutions, publishers, or journals, might exert an influence over an article's later impact through the accompanying press releases they issue. It should be noted that such influence would obviously only constitute one component of a multifaceted and highly complex mélange of factors that affect research articles' impact metrics (Tahamtan et al., 2016).

For many research institutions and scholarly publishers, press releases constitute the quintessential instrument for marketing new knowledge (Autzen, 2014). As Carver (2014, p. 2) describes it, the press release is "essentially a short news article written in a journalistic style that explains a newly published scientific result in a common and not too specialized language". The arguably most important international platform for the dissemination of press releases on science is *EurekAlert!*, set up by the *American Association for the Advancement of Science* in 1996. According to Vrieze (2018), with over 5,000 active public information officers and more than 14,000 registered journalists from over 90 countries, the platform has become for scientific press and news releases "what *Google* is for searching and *Amazon* for online shopping".

Methods & Data

We use press release data provided directly by *EurekAlert!*. The data contains a comprehensive list of 11,110 unique DOIs of research publications for which at least one press release was published on *EurekAlert!* in 2016, 10,859 of which refer to journal articles according to the *Crossref* REST API. To be able to identify the articles with comparatively high and those with comparatively low metrics among the set, we obtain altmetric data for the 10,859 journal article DOIs promoted on *EurekAlert!* in 2016 from *Altmetric.com* as well as citation counts from the CCB databases[1] via the respective services' APIs. Metrics data was retrieved in October 2020. As a starting point for this research in progress, we focus on articles with particularly high or low impact regarding the six types of indicators which Lemke (2020) found to be associated with press release promotion: *Web of Science* citation counts, tweet mentions, *Facebook* mentions, blog mentions, mainstream media mentions, and *Mendeley* readership counts.

Using the method of characteristic scores and scales introduced by Glänzel and Schubert (1988), out of our 10,859 journal article DOIs we extract 2 subsets of articles for each of these six indicators: one set of articles performing remarkably and one set of articles performing poorly regarding said indicator. This provides us with 12 subsets of articles. From each of these 12 sets we draw 10 random articles, for which we then retrieve the respective press releases that promoted them from *EurekAlert!*. This leads to a set of 124 press releases to analyze.

The press releases are then coded manually by two coders with an inductive approach, meaning the iterative coding of subsets to continuously develop a codebook of properties that might be helpful in the explanation of different press releases' varying promotional effects for the featured research articles. Starting points for which structural and content-related characteristics to expect from press releases to research articles exist in the form of guidelines on how to write them.[2] Each round of coding consists of thorough reading and simultaneous notetaking. These notes are iteratively reviewed by both coders to identify relevant properties and ranges of values these properties can take. The revised codebook is then again applied to the dataset.

Preliminary Results & Discussion

In the following subsections, we first briefly describe our dataset and then summarize preliminary findings regarding properties in which the coded press releases differ from each other, which might also affect their (and promoted research articles') later uptake.

Article properties

In total, the 120 articles whose 124 press releases we analyzed were published in 72 different journals from 21 publishers, *Science* and *Nature* being the most strongly represented journals (with 14/12 articles respectively). We manually identified the articles' disciplines by reading their press releases (which should enable more accurate article-level mappings than consulting publishing journals' *Web of Science* subject categories), finding most of them to be related to medicine (56 press releases) or biology (28 press releases), followed by psychology, chemistry, economics, archeology, and geology. The dominance of life science topics, both in public media's science coverage in general as well as on *EurekAlert!* in particular, is in line with findings from previous studies (e.g., Elmer, Badenschier & Wormer, 2008; Hahn & Lemke,

[1] CCB refers to the German *Competence Centre for Bibliometrics*, an institution hosting annually updated citation databases built on data from *Web of Science*. The citation data used in this study reflects a state from April 2020.

[2] For examples, see https://www.cbsnews.com/news/how-to-write-a-press-release-with-examples/, https://esahubble.org/about_us/scientist_guidelines/, https://www.asbmb.org/education/science-outreach/how-to-write-a-press-release, or https://service.prweb.com/resources/article/editorial-guidelines/.

2020). The majority (116) of the 120 journal articles were published in the same year as their respective press release, 2016, two were published towards the end of 2015, the remaining two early in 2017. Table 1 describes the 60 remarkably performing and the 60 poorly performing articles regarding the six article-level indicators considered in this study, as well as their publishing journals' impact factors in 2016.

Of the 124 associated press releases, 83 (46 for high, 37 for low performing articles) had been submitted to *EurekAlert!* by institutes conducting research, e.g., universities, laboratories, or hospitals, 41 (17 for high, 24 for low performing articles) by scholarly publishers.

Table 1. Indicator-related statistics of both article groups

	Poorly perf. articles			Remarkably perf. articles		
	Mean	Median	SD	Mean	Median	SD
Citations	22.1	15	23.9	136.2	127	118.7
Mainstream media mentions	12.0	8	14.5	68.9	52.5	63.9
Blog mentions	1.2	1	1.8	11.1	10	8.5
Tweet mentions	17.5	8	24.5	282.2	179	291.1
Facebook mentions	1.2	1	1.7	14.4	10.5	17.2
Mendeley readership counts	79.0	54	89.5	450.7	386	408.9
Journal Impact Factor (2016)	8.337	4.259	9.948	24.130	16.761	14.110

Structure of press releases

Press release texts were analyzed for whether they follow a clearly discernible structure that supports a comprehensible line of argumentation. Ultimately, the coders evaluated a press release's structure as *clear*, *slightly unclear*, or *unclear*. A *clear* buildup could for instance start with a concise problem statement and a brief snapshot of the research, followed by definite paragraphs with comprehensible individual functions, e.g., more detailed descriptions of the research's methods, added value, and further implications. An *unclear* structure on the other hand might merge several diverse parts of information within few, long paragraphs. Another indicator for a structure not being *clear* might be the existence of unnecessary repetitions. Overall, most (96) of the press releases were evaluated as having a clear structure. Another 17 press releases stood out as having an unclear structure, for instance because of sudden jumps in their lines of argumentation or the connection between problem statement and the reported results being vague. Only 5 of those 17 press releases belonged to articles from the remarkably performing group, while the remaining 12 promoted articles which performed poorly.

Accessibility of press releases

Another aspect in which coded press releases differed is their accessibility due to their linguistic or technical complexity. Coders differentiated between *good*, *medium* and *bad* accessibility, depending on whether they deemed the press release understandable for readers that are no experts in the related field of research. A *bad* accessibility could for example result from the press release containing a high number of unexplained technical terms that an average reader would likely have to look up. Most (100) press releases were found to have *good* accessibility. Only 13 press releases were evaluated as having *bad* accessibility, most often because they were written in a highly technical language without sufficient explanations. In this case, 9 of these 13 belonged to articles from the indicator-wise poorly performing group of articles.

Engaging narrative

As another step, coders assessed the press releases' use of engaging narratives to report findings. Primarily, this refers to the press release's author's writing style and not to the promoted research's topic. We do however note that as certain topics will be more suitable for an engaging or emotional style, this category will be affected more significantly by the promoted article's content than the press release´s structure or linguistic accessibility. Coders assessed the degree to which a press release contained an emotionally engaging style on a five-level Likert scale from *low* (1) over *medium* (3) to *high* (5). *High* emotionality can be the result of a narrative that creates tension, or one that particularly effectively depicts research findings' relevance. *Low* emotionality on the other hand is indicated by an austere writing style dominated by technical information. Most press releases were found to have a *low* (58) or *low-medium* (21) level of engaging narratives, a *medium* score was assigned 19 times, *medium-high* 12 times, and *high* 14 times. Across all scores the two article groups were represented in almost perfectly equal shares. Thus, the degrees to which press releases were found to vary regarding their use of narratives does not seem to correlate with articles' performance concerning metrics.

Our finding of comparatively large shares of poorly performing articles among those with badly structured or linguistically inaccessible press releases suggests that there might be some association between an article's metrics performance and certain qualities of its press releases. However, we need to keep in mind two limitations of our preliminary results: first, our approach (for now) suffers from a small sample size, hindering any observations' generalizability. Second, we cannot make statements about potential associations' directionalities yet. Just as high metrics could (partially) result from the visibility generated by well-made press texts, bad press texts could (partially) be the result of 'bad source material'. As hinted at earlier, the ease with which an engaging narrative can be found to report about an article's findings certainly depends on its topic, and even press releases' structural or linguistic properties might not be independent from the promoted article's content. Thus, the latter's inherent flaws could at the same time be part of the reason for the article receiving lower metrics, as well as for its press releases to be more likely to be perceived as inaccessible or poorly structured.

Furthermore, the quality of press releases could be strongly connected to the publishing journal or institution, just like metrics are strongly affected by the journal an article is published in. Perhaps part of particularly prominent journals' success can be explained by their superior PR, which produces more accessible and better structured press texts than their competitors do. These are questions we aim to tackle in more detail in our research project's subsequent steps.

Conclusion & Outlook

First results from our coding revealed several inherent properties regarding which press releases to scholarly articles differ from each other. Also, the quantitative findings may suggest associations between some of these factors and a promoted article's later metrics. This research is still in an early stage and follow-up studies could take numerous directions.

The next steps in this study will consist of consolidating the observations made during coding to a reusable codebook, its validation through its application by 'blind' raters, which were not involved in its conception, and its subsequent formal assessment in terms of inter-rater-reliability. Regarding further manual coding, it could be insightful to additionally assess press releases regarding their extent of research content, i.e., the degree of detail with which the research itself is presented in them. It might also be interesting to complement raters' judgments of press releases' accessibility with automatically calculated measures of readability, e.g., Coleman-Liau index. Furthermore, upcoming efforts will go into analyzing correlations between properties' manifestations, individual types of metrics, and external factors like the press release's publisher or the promoted article's journal. Moreover, it could be worthwhile to retrieve full texts of news media reporting on the articles from our sample as well as other research articles that cited them to examine the whole alleged chain *research article – press release – media coverage – citation*. Such a comprehensive look might shed light on how certain qualities are carried over (or lost) between different formats and help to better explain the mechanisms behind the association between press coverage and increased article impact.

In later steps, machine learning methods could be applied to make the coding scalable for larger samples of press releases. Another in-depth content analysis could investigate how homogeneous press releases behave across individual journals or publishers – such a study could both enhance the effectiveness of a machine learning-based application of the codebook, as well as provide insights into how individual promotional activities of major players in the world of scholarly publishing shape external science communication.

Acknowledgments

This study is an outcome of research project MeWiKo, funded by the German Federal Ministry of Education and Research (grant number 01PU17018A). We thank *EurekAlert!* and the CCB for providing our project with access to their data.

References

Autzen, C. (2014). Press releases — The new trend in science communication. *Journal of Science Communication, 13*(03). doi:10.22323/2.13030302

Badenschier, F., & Wormer, H. (2012). Issue Selection in Science Journalism: Towards a Special Theory of News Values for Science News? In S. Rödder, M. Franzen, & P. Weingart (Eds.), *The Sciences' Media Connection –Public Communication and its Repercussions* (pp. 59–85). Dordrecht: Springer Netherlands. doi:10.1007/978-94-007-2085-5_4

Carver, R. B. (2014). Public communication from research institutes: Is it science communication or public relations? *Journal of Science Communication, 13*(3), C01. doi:10.22323/2.13030301

Chapman, S., Nguyen, T. N., & White, C. (2007). Press-released papers are more downloaded and cited. *Tobacco Control, 16*(1), 71–71. doi:10.1136/tc.2006.019034

Elmer, C., Badenschier, F., & Wormer, H. (2008). Science for Everybody? How the Coverage of Research Issues in German Newspapers Has Increased Dramatically. *Journalism & Mass Communication Quarterly, 85*(4), 878–893. doi:10.1177/107769900808500410

Fanelli, D. (2013). Any publicity is better than none: Newspaper coverage increases citations, in the UK more than in Italy. *Scientometrics, 95*(3), 1167–1177. doi:10.1007/s11192-012-0925-0

Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Information Science, 14*(2), 123–127. doi:10.1177/016555158801400208

Hahn, O., & Lemke, S. (2020). An Exploration of Scientific Press Releases in the Context of Altmetrics. Presented at the *altmetrics20 Workshop*. doi:10.5281/zenodo.4446908

Kiernan, V. (2003). Diffusion of News about Research. *Science Communication, 25*(1), 3–13. doi:10.1177/1075547003255297

Lemke, S., Mehrazar, M., Mazarakis, A., & Peters, I. (2019). "When You Use Social Media You Are Not Working": Barriers for the Use of Metrics in Social Sciences. *Frontiers in Research Metrics and Analytics, 3*, 39. doi:10.3389/frma.2018.00039

Lemke, S. (2020). The Effect of Press Releases on Promoted Articles' Citations and Altmetrics. Presented at the *Metrics 2020: ASIS&T Virtual Workshop on Informetrics and Scientometrics Research*. doi:10.5281/zenodo.4351360

Nicholas, D., Herman, E., Jamali, H. R., Abrizah, A., Boukacem-Zeghmouri, C., Xu, J., Rodríguez-Bravo, B., Watkinson, A., Polezhaeva, T., & Świgon, M. (2020). Millennial researchers in a metric-driven scholarly world: An international study. *Research Evaluation, 29*(3), 263–274. doi:10.1093/reseval/rvaa004

Phillips, D. P., Kanter, E. J., Bednarczyk, B., & Tastad, P. L. (1991). Importance of the Lay Press in the Transmission of Medical Knowledge to the Scientific Community. *New England Journal of Medicine, 325*(16), 1180–1183. doi:10.1056/NEJM199110173251620

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics, 107*(3), 1195–1225. doi:10.1007/s11192-016-1889-2

Vrieze, J. de. (2018, September 27). EurekAlert! Has spoiled science news. Here's how we can fix it. Retrieved January 11, 2021, from VWN - Vereniging voor Wetenschapsjournalistiek en -communicatie Nederland website: https://www.vwn.nu/?p=2225/

Citing a retracted paper: the case of Wakefield's article that correlates vaccine and autism

Jacqueline Leta[1], Kizi Araujo[2] and Stephanie Treiber[3]

[1] *jleta@bioqmed.ufrj.br*
Federal University of Rio de Janeiro, Rio de Janeiro (Brazil)

[2] *kiziaraujo@yahoo.com.br*
Oswaldo Cruz Foundation, Rio de Janeiro (Brazil)

[3] stephanie.treiber@bioqmed.ufrj.br
Federal University of Rio de Janeiro, Rio de Janeiro (Brazil)

Abstract

The present study aims to describe and discuss some characteristics of the citing documents of Wakefield's paper, a retracted article due to fraudulent data and analysis on the relationship between vaccine and autism. All metadata of the 1,578 citing documents of Wakefield's paper were downloaded from Scopus, considering two periods before (1998-2004) and post retraction (2005-2020). The main results indicate that the 79% of the citing documents are post retraction. In this period, we found the predominance of articles as well as other documents written in English but also in a wider-ranging of idioms. These documents are mostly associated with medicine and social sciences. These findings are a clear indication that this retracted paper is still popular and dispersed within a broader spectrum of fields in the global arena, from medicine and health sciences to social sciences. For future analysis, we intend to analyze the citing documents qualitatively, seeking to better understand the motivations for citing Wakefield's paper even after its retraction.

Introduction

During the 20th century, scientific journals have increased considerably not only in terms of quantity but also in field coverage. The success of this means of communication is related to many reasons, such as a lower production cost and logistics complexity as well as a faster speed in the diffusion of the new knowledge. Nevertheless, its main distinction concerns the peer review process, which acts as a self-control process to guarantee the quality of scientific work. (Meadows, 1974)

The process of peer review was introduced in the 1600's by Henry Oldenburg, the editorship of The Philosophical Transactions of the Royal Society that is considered one of the first world scientific journals. Gradually, this practice spread and was incorporated into the routine of the scientific journals' publishing process. At first, the main editor and an in-house staff of specialists assumed the responsibility for reviewing the submitted works. After the 1950's, experts out of the staff started to be invited to review and approve (or reject) manuscripts, a model that persists up to the present. (Burnham, 1990)

Although the peer review process has been consolidated as a central step in the publishing process and it has been supported by the majority of the world scientific community, the integrity of this process has been frequently questioned and criticized. Bornmann's review of literature on peer review presents a comprehensive view on the main issues raised by defenders and critics of peer review process (Bornmann, 2011)

Additionally, the peer review process has not been able to detect or to avoid the increase in science misconduct during the last decades, as estimated by the growing number of retractions (Fang, Grant Steen & Casadevall, 2012; Bar-Ilan & Halevi, 2017). It is true that the quality and the confidence of peer review has changed a lot due to the "high volume of submissions, turnover of editors, and dependence upon ad hoc reviewers and often a part-time (non-professional) managing editor" (Fox, 1994).

Such new context partially explains the increase in misconducting, while most are related to scientists' conducts. Yet, according to Steneck (2006), since 1980's the debate on misconduct in the scientific environment led to the consensus that misconduct is associated with two main groups of irresponsible research behaviors, named as (a) questionable practices, that is, when a study presents misrepresented, inaccurate or biased data or analysis and (b) deliberate misconduct practices that include cases of fabrication, falsification, and plagiarism.

An example of serious misconduct is the study developed by Andrew Wakefield and published in 1998 in *The Lancet* journal (Wakefield *et al.*, 1998), one of the world's most prestigious journals. The paper suggests a relationship between the triple viral vaccine (which protects against measles, rubella and mumps) and the development of autistic behavior in children. But in 2004, Wakefield's article was proved to be fraudulent and was retracted.

Despite the retraction, the myth of vaccines causing autism still persists in different regions of the world, since Wakefield's paper can be easily found in the internet and different social media. In fact, a quick look at the platform PlumX Metrics, owned by Scopus, revealed that up to January 2021 there were more than 9,200 shares, likes or comments of this paper on Facebook and almost 2,000 tweets. Similarly, in Google Scholar, an academic media, this paper is also alive: there were more than 3,600 documents citing it; one third of these documents were published in 2015 or after.

Due to its widespread and current repercussion, some studies consider this article as a central element in the resurgence and strengthening of the global anti-vaccine movement (Hussain, 2018). Hence, considering the social impact of Wakefield's fraudulent article and its long-trajectory within the scientific environment, the present study aims to find out answers to a central research question: why is this article still cited? In order to start getting evidence and answers for this question, we investigate some general characteristics of Wakefield's citing documents before and post retraction.

Methodology

In January 2021, Wakefield's paper (Wakefield *et al.*, 1998) was searched in Scopus database, a multidisciplinary specialized database, which is available for all Brazilian research institutes and universities through Portal Periódicos (http://www-periodicos-capes-gov-br.ezl.periodicos.capes.gov.br/index.php?). The search result display paper contained 1,578 citing documents which were downloaded. Among the citing documents, there was one published in 2021, but it was not included in the analysis.

Metadata were downloaded in two files according to the year of publication: 1998-2004 (named before retraction) and 2005-2020 (post retraction). Both files were downloaded in a CSV format and converted to Microsoft Excel format in order to proceed the descriptive analysis.

Results

The results are presented in three sessions. The first session aims to show the number of citing documents along the studied period and their main typology. The second session focuses on the origin and idiom of correspondence authors of these documents while the last one is about their main thematic. In all sessions, results are presented in two periods: before and post retraction of Wakefield's paper in 2004.

Time trend and typology

Analyzing the year distribution of Wakefield's paper citing documents (Figure 2), we observed a great number of citing documents just after its publication (more than 20 citations are in the first year after publication), with a very expressive growth between the years 2000 and 2003. In 2005, a year after the paper's retraction, a slight drop in the number of citing documents was noted. However, the downward trend is not maintained and Wakefield's paper remains highly cited throughout the analyzed period, including growth peaks in the years 2010 and 2013.

It is noteworthy that 79% of the citing documents (n = 1,248) occur post retraction of Wakefield's article (2004). In this period, we found an average of 78 citing documents per year, while the period before retraction presents an average of 47 citing documents per year. This picture clearly indicates the strong repercussion and impact of Wakefield's article within the scientific community since its publication, but mainly, post retraction.

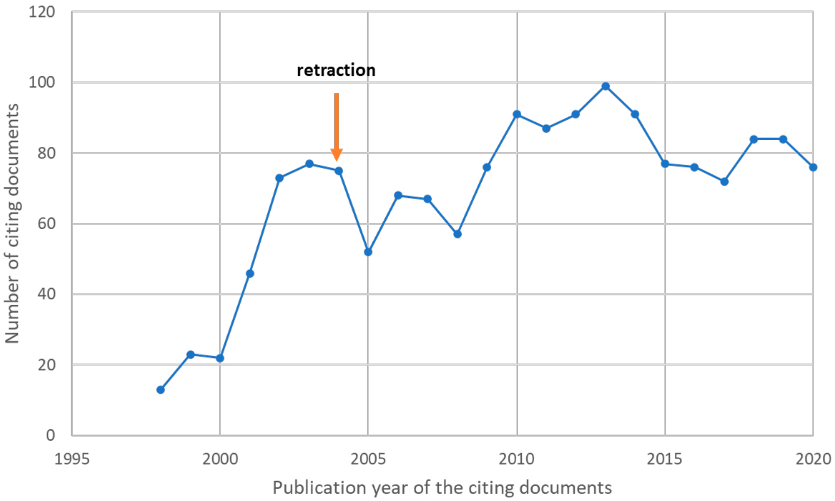


Figure 2. Time trend of citing documents of Wakefield's paper. Scopus

Regarding the typology of the citing documents (Table 1), we identified that most are of the article typology. Comparing before and after retraction, we noted that the share of citing documents in article typology increased 5%, suggesting that a larger number of the new scientific knowledge is still grounded on a retracted paper.

Considering the other typologies, we noted that a significant fraction of the citing documents post retraction comes from books (10.9%) and book chapters (5.6%); in the period before retraction, the share of both typologies was around 1-2 %. A quick look at the list of citing documents classified as books and chapters reveals that they are distributed in a broad spectrum of specialties in the health field, especially pediatrics, and there are also titles in other more general fields, such as scientific integrity. Such a general profile is somehow expectable, since the vaccine mentioned in Wakefield's paper is administered to children, so it is an issue that interests pediatricians, while the errors and misconduct attributed to this paper may serve as a typical case of misconduct in science.

Table 1. Citing documents of Wakefield's paper according to typology, before and post retraction in 2004. Scopus

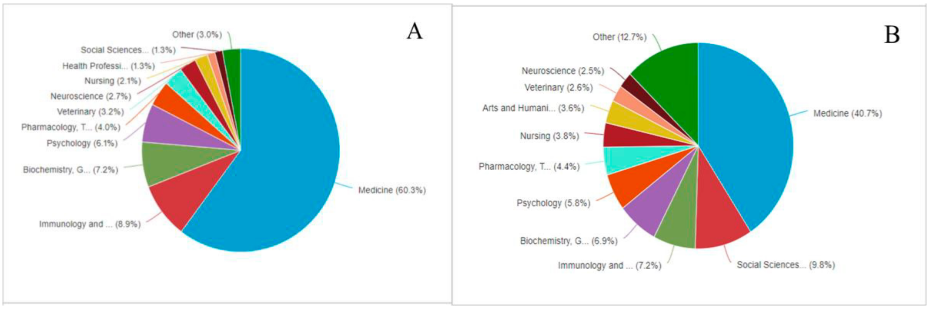
Typology	*Before*		*After*	
	Number	%	Number	**%**
Article	139	42.2	591	47.4
Review	107	32.5	268	21.5
Book Chapter	6	1.8	136	10.9
Editorial	23	7.0	79	6.3
Book	3	0.9	70	5.6
Note	15	4.6	40	3.2
Letter	9	2.7	28	2.2
Conference Paper	17	5.2	19	1.5
Short Survey	10	3.0	16	1.3
Erratum		-	1	0.1
Total	**329**		**1,248**	

Areas and thematic

Figure 4 presents the areas of the citing documents before (Figure 4A) and post retraction (Figure 4B). We noted that the general profile of the areas of the citing documents remains similar in both periods.

However, post retraction (Figure 4B), there was a strong reduction in the percentage of citing documents in the area of medicine (from 60% to almost 41%) and a remarkable increase in the percentage of documents in the area of social sciences (from 1.3% to almost 10%). Such increase may be related to ethical issues in science, a common and growing thematic among researchers from social sciences.

Figure 4. Area of citing documents of Wakefield's paper before (A) and post (B) retraction. Scopus



However, post retraction (Figure 4B), there was a strong reduction in the percentage of citing documents in the area of medicine (from 60% to almost 41%) and a remarkable increase in the percentage of documents in the area of social sciences (from 1.3% to almost 10%). Such increase may be related to ethical issues in science, a common and growing thematic among researchers from social sciences.

Conclusion and some remarks

The present study reveals some general characteristics of citing documents of Wakefield's paper that, despite being retracted in 2004, continued to be cited by the scientific community. We found that 79% of citing documents were published post retraction (Figure 2). In this period, most of the citing documents were in the format of articles (Table 1), written in English but also in a wide-ranging of idioms and countries (data not shown). At last, we showed that the citing documents after the retraction of Wakefield's paper are still mostly associated with medicine but we also noted an increase in documents related to social sciences (Figure 4).

These findings indicate that this fraudulent paper is still popular and dispersed within a broader spectrum of fields in the global arena, from medicine and health sciences to social sciences. Other analysis, not shown in the present paper, reveal that post retraction includes a large number of citing documents related to issues on science communication and ethics in scientific research as well as science education and parental responsibilities.

It is important to highlight that Wakefield's paper appeared as the number one in the ranking of most cited ranked articles, all published in very prestigious journals, as described by Fang, Steen & Casadevall (2021). The authors have found that retraction can cause a continuous reduction in the number of citations and, in some cases, also "an immediate and severe decline in citations". This conclusion, however, is not in accordance with the study developed by Candal-Pedreira et al (2020). These authors have found no association between retraction and citations in the long term.

In order to better understand the reasons beyond the persistent popularity of retracted papers, Bar-Ilan et al (2017) investigated 237 citing documents of 15 retracted articles found in ScienceDirect. After identifying the context of each citation (positive, negative or neutral), authors conclude that most of citing documents were used in a positive context, even in cases where retraction was due to serious and deliberate misconduct, as fabrication and falsification data. Authors highlight that the easy access to the retracted paper on the internet and the intensive media attention on retracted papers (such as Wakefield's paper) may contribute to continuing (and sometimes growing) interest in these studies not only by the general public but also by scientists, who may cite them in a positive or negative context.

The present study indicates a change in the thematic profile in the post retraction period, more focused on social science issues, which may be an indicative of a reflective citation, probably a negative one, on Wakefield's paper. However, the set of data and results presented here do not allow us to make such a statement. For future analysis, we intend to analyze the citing documents qualitatively, seeking to better understand the motivations for citing Wakefield's paper even after its retraction. These results can be useful for a better understanding of the scenario of scientific communication during the pandemic, in which we are witnessing a rapid repercussion of papers that in a short time have failed to confirm their main conclusion.

Acknowledgments

This work has financial support from CNPq through the Doctoral scholarship granted to Stephanie Treiber and the funding awarded to the research project n. 434.146/2018-8.

References

Bar-Ilan, J. & Gali, H. (2017). Post retraction citations in context: a case study. *Scientometrics* 113 (1) 547-565.

Bornmann, L. (2011). Scientific peer review. *Annual review of information science and technology* 45 (1) 197-245.

Burnham, J. (1990). The Evolution of Peer Review. *Journal of American Medical Association*, 263, 1323-29.

Candal-Pedreira, C. et al. (2020). Does retraction after misconduct have an impact on citations? A pre–post study. *BMJ Global Health* 5 (11) 1-7.

Fang, F.C., Grant Steen, R. & Casadevall, A. (2012) Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences* 109 (42) 17028-17033.

Fox, M.F. (1994) Scientific misconduct and editorial and peer review processes. *The Journal of Higher Education* 65 (3) 298-309.

Hussain, A. et al. (2018). The anti-vaccination movement: a regression in modern medicine. *Cureus* 10 (7) 1-8.

Meadows, A.J. (1974). *Communication in Science*. Butterworths: Seven Oaks.

Steneck, N.H. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and engineering ethics* 12 (1) 53-74.

Wakefield, A.J, et al. (1998). Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet*. 351 (9103) 637-41.

Have the European Union programmes made a difference to biomedical research outputs?

Grant Lewison[1]

[1] *grantlewison@aol.co.uk*
King's College London, Department of Cancer Policy, Guy's Hospital, Great Maze Pond, London SE1 9RT, (UK)

Abstract

This study was designed to see if the several European Union (EU) research programmes, including the Coöperation in Science and Technology (COST) activities, had made a measurable difference to the Member States (MS) who took part in them. We looked at the internationally collaborative biomedical research outputs of six MS that joined the EU at different times (Greece, Spain, Sweden, Poland, Romania and Croatia), and compared them with those of selected similar non-EU countries (Egypt, Brazil, Norway, Russia, Ukraine, Serbia) to see if there was more work done jointly with other EU MS compared with the Rest of the World. We also checked whether the citation impact of their purely domestic biomedical research had increased more rapidly than that of the six comparator countries. The results were somewhat mixed, probably because EU support only accounted for a small fraction of the total cost of this research, but there were some discernible benefits. Suggestions are made for future investigations that would be more likely to reveal the effects of EU programmes.

Introduction

Research and development has been a major activity of the European Union (EU), and of the European Communities (EC) before that. In fact, each of the three Communities (the European Coal and Steel Community (founded in 1951), the European Atomic Energy Community (EURATOM), and the European Economic Community (both founded in 1957)) had their own research programmes. The latter grew to become much the largest, and the successive Framework Programmes, followed by Horizon 2020 and Horizon Europe, have become a major source of both funding and collaboration within the European Union, and in 13 other Associated Countries (including Israel, Norway, and Switzerland) who participate in the same way as Member States (MS). Their governments pay into the programme budgets on the basis of their national wealth.

Despite the current size of the recent EU programmes, they represent only a small fraction of the total R&D expenditure by the MS. The evaluation literature covers both the results of the EU research grants and other coördination activities such as Co-Operation in Science and Technology (COST), and studies on overall European research outputs in specific subjects. In the 20th century, the main concern was European success relative to that of the other members of the triad (the USA and Japan; Tikal, 1993; Leydesdorff, 2000), but the focus has now shifted to comparisons with the USA and China (Leydesdorff, Wagner & Bornmann, 2014; Evangelatos et al., 2018). Several papers examined the outputs of individual EU-funded research programmes (de Nettancourt & Klepsch, 1990; Lewison, 1994; Weydert, 2000; Sauter et al., 2011; David, 2016). However, rather more papers were concerned with the distribution of publication outputs within the EU, particularly papers on medical research (Parodi et al., 1993; Mela & Cimmino, 1998; Hefler, Tempfer & Kainz, 1999; Ramos et al., 2004; Soteriades & Falagas, 2005; Begum et al., 2016; Begum et al., 2017a; Begum et al., 2018; Kalo et al., 2019; Pallari et al., 2020).

This paper seeks to determine the overall effects of EU biomedical research activities by means of two comparisons between the outputs of six EU MS and those of six comparable countries that were not MS, although some of them were associated with the EU's programmes. The first parameter was the amount of collaboration between each of the pairs of countries and (a) other EU MS (at the time the new MS joined the EU) and (b) the rest of the world (RoW). The second parameter was the five-year mean citation score of *domestic* research papers. These were distinguished from ones with international co-authorship because their citation impact would be strongly influenced by the foreign co-authors. The idea was to see if EU Membership had led to an improvement in the impact of the country's own researchers. We compared the situation in the years *before* each of the six selected MS joined the EU, and in the subsequent years.

The six EU MS and the six comparator countries are listed in Table 1, with their International Standards Organization digraph (ISO2) codes, and their Gross Domestic Products in the accession year and in 2015 to show that they were broadly comparable. However, there have been great changes in the GDP of some countries, notably Egypt, Croatia and Ukraine, but very little in Romania.

Table 1. List of six European Union Member States (EU MS), with their ISO digraphs, the dates that they joined the EU (or the EC), and their GDPs both then and in 2015 (billion US $). Similar data are given for six comparator countries.

EU MS	*ISO2*	*Accession*	*GDP, US $ bn*		*Comparator*	*ISO2*	*GDP, US $ bn*	
			Then	*2015*			*Then*	*2015*
Greece	GR	1981	52	195	Egypt	EG	23	331
Spain	ES	1986	251	1193	Brazil	BR	268	1804
Sweden	SE	1995	264	496	Norway	NO	152	387
Poland	PL	2004	255	477	Russia	RU	591	1331
Romania	RO	2007	172	178	Ukraine	UA	111	906
Croatia	HR	2013	58	487	Serbia	RS	46	372

All of the comparator countries have participated in EU programmes to some extent. For example, Norway has been an Associated country since 1995, when it declined to accede to the EC as a result of its referendum. Poland and Romania have also been involved in EU programmes since the early 1990s when the EU envisaged enlargement to the countries of Eastern Europe, following the fall of the Berlin Wall in June 1990, although they were less active than other Eastern European MS, perhaps because of a lack of domestic investment in R&D. There have also been some European support programmes with Russia, Ukraine and Serbia. Egypt and Brazil also took some part in EC and EU programmes, though not as intensively as Member States. Finally, most of the comparator countries have been involved in other trans-national partnerships, such as the Centre d'Etudes de Recherche Nucléaire (CERN) in Geneva and the European Molecular Biology Laboratory (EMBL) in Heidelberg.

Methodology

For the 12 countries listed in Table 1, we determined their biomedical research outputs in each year from 1975 to 2019 in the Web of Science (WoS, © Clarivate Analytics) by means of a complex filter based on address words or contractions (Lewison & Paraje, 2004). These included names of diseases or parts of the body, such as *AIDS, CANC, HEART*, names of pharma companies, such as *BAYER, GLAXO**, *LILLY,* organisations supporting biomedical research, such as *INSERM**, *MRC, NIH,* and places where such research was conducted, such as *BETHESDA, HOSP*, INFIRM, KAROLINSKA**. [For the bottom three countries in each list, only papers from 1995 were identified.]

In order to check that the results would not be biased by some of the 12 selected countries having a rather unusual distribution of diseases within their biomedical research portfolios, we also determined their outputs in cardiovascular research and in cancer research, the two leading causes of death in high- and middle-income countries. These outputs correlated very closely with the biomedical research outputs ($r^2 = 0.97$ and $r^2 = 0.92$, respectively), so we felt confident that the latter were representative of health research in the selected countries.

We also determined the number of papers in collaboration only with the other EU MS at the time that they joined, and only with the RoW, and with no international collaboration. These four numbers of papers for each year were used to calculate the additions to their domestic outputs from the EU MS, from the RoW, and from both.

We then determined the mean five-year citation counts (Actual Citation Impact, ACI) for the domestic papers for each of the 12 countries for each fifth year (1975, 1980, etc). For a few countries in the most recent two years (2010, 2015) the numbers of papers exceeded the limit of 10,000 for the determination of citation counts in the WoS. In these instances, we divided the set of papers into two on the basis of the initial letters of their journals, thus: SO=(A* or B* or C* ....), and added the resulting citation numbers.

The results have mainly been presented graphically, with two parallel graphs for each parameter, one for the top three pairs of countries and dates 1975-2019, and the other for the bottom three pairs and dates 1995-2019. Some sample graphs showing the outputs of domestic biomedical papers from one country, Sweden, and the additions from the EU MS, the RoW, and for both have also been presented.

Now the world-wide output of biomedical research papers in the WoS has increased greatly over the 45-year study period, by a factor of almost eight, and the journal coverage has increased by a factor of more than six, see Figure 1. The number of journals covered shot up by over 4800 (46%) in 2015 when the Emerging Sciences Citation Index was created and incorporated in the WoS. Many of these were published in non-traditional countries. Another reason for the rise in world biomedical research output was the rapidly increasing presence of China, which went up from 0.13% in 1975 (171 papers) to over 19% in 2019 (196,843 papers). As a result, it may be more revealing to show the outputs of these European (and other) countries as percentages of world output in the given years.

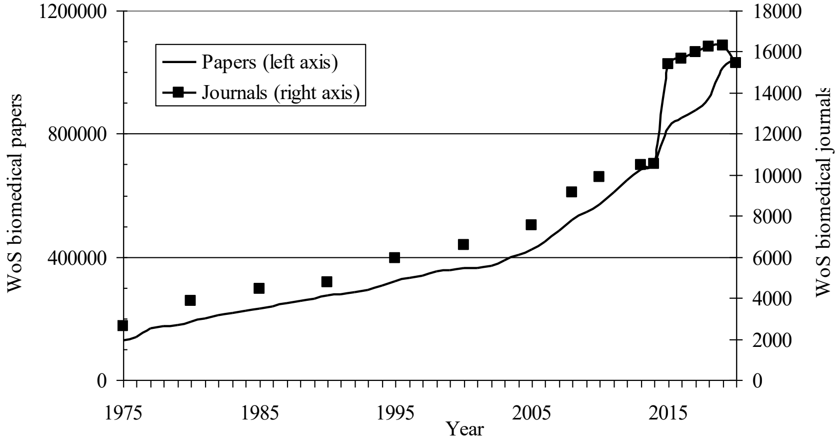


Figure 1. World outputs of biomedical research papers in the Web of Science (as defined by the address-based filter), 1975-2020, and the numbers of journals with one or more papers in the data-set. Note the sharp increase in the numbers of journals in 2015 because of the advent of the Emerging Sciences Citation Index.

Results

Numbers of papers and collaboration

Most European countries have increased their biomedical research outputs over the study period, but their percentage presence in the world has declined, mainly because of the very rapid rise in the output of China. Figures 2 and 3 show the absolute and percentage values respectively for one country (Sweden, which joined the EC in January 1995), with the outputs for its domestic papers, and the additions from other EC Member States, from the Rest of the World, and from both. The parameter of interest is the increase in the ratio of numbers of extra papers with collaboration from other EU MS (*i.e.*, the sum of the first and third additions) to the numbers of domestic papers. This is shown in Figure 4 for the first three selected MS to join the EC (Greece, Spain, and Sweden) and their comparators (Egypt, Brazil, and Norway). Figure 5 shows the corresponding ratios for the remaining three MS (Poland, Romania, and Croatia) and their comparators (the USSR/Russia, Ukraine, and Serbia).

Table 2. The ratio of the additional papers in collaboration with other EU MS to domestic papers for six new MS and their comparator countries

	Accession year	*Plus 5 years*	*Plus 10 years*
Greece/Egypt	0.82	1.03	1.25
Spain/Brazil	0.67	0.78	0.94
Sweden/Norway	0.65	0.81	0.77
Poland/USSR-Russia	0.78	0.85	0.61
Romania/Ukraine	0.51	0.26	0.82
Croatia/Serbia	1.20	1.33	

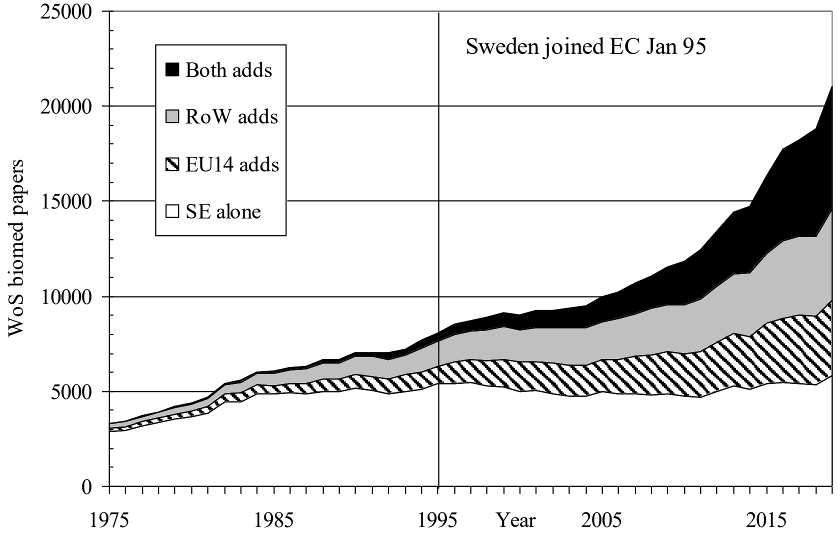


Figure 2. Biomedical research outputs for Sweden in the Web of Science, 1975-2019. Domestic papers alone, the addition of papers co-authored with other EU MS only, with countries from the Rest of the World only, and with both. Absolute numbers of papers.

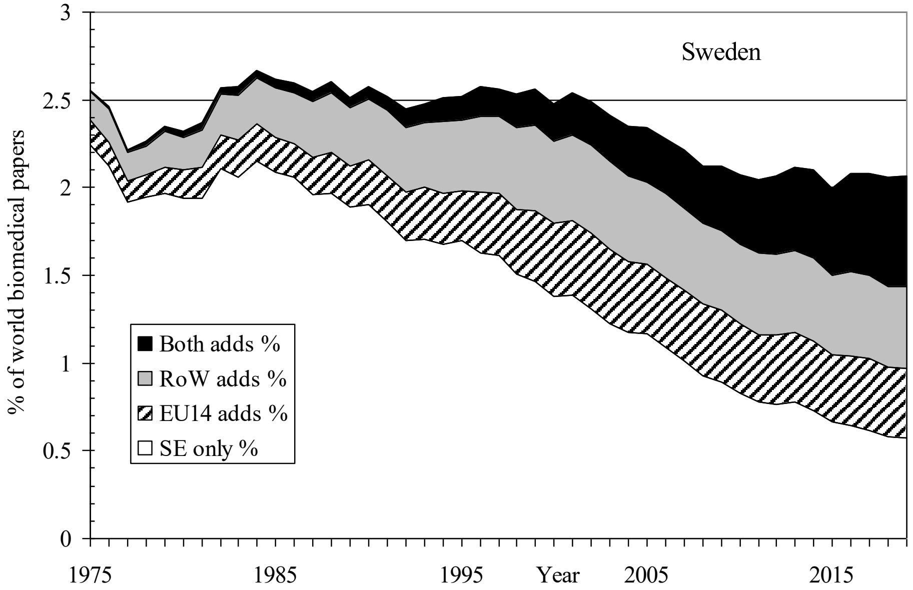


Figure 3. Biomedical research outputs for Sweden in the Web of Science, 1975-2019. Domestic papers alone, the addition of papers co-authored with other EU MS only, with countries from the Rest of the World only, and with both. Percentages of world outputs.

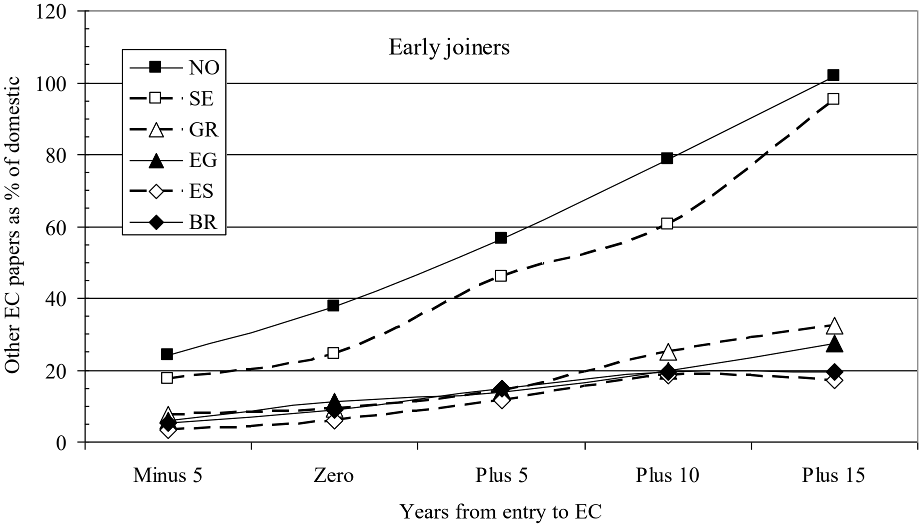


Figure 4. The changing ratio of additional papers co-authored with another EC Member State to domestic papers for three pairs of early joining countries

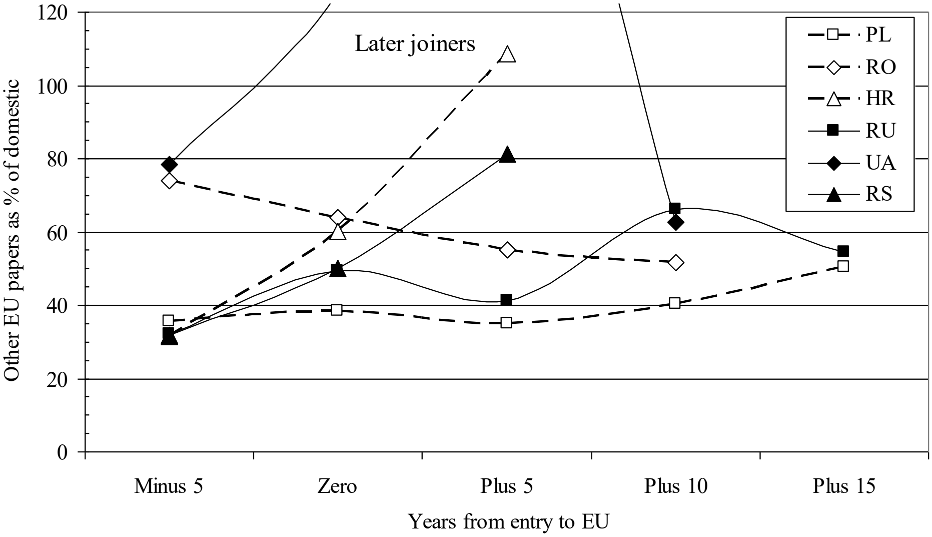


Figure 5. The changing ratio of additional papers co-authored with another EC Member State to domestic papers for three pairs of later joining countries

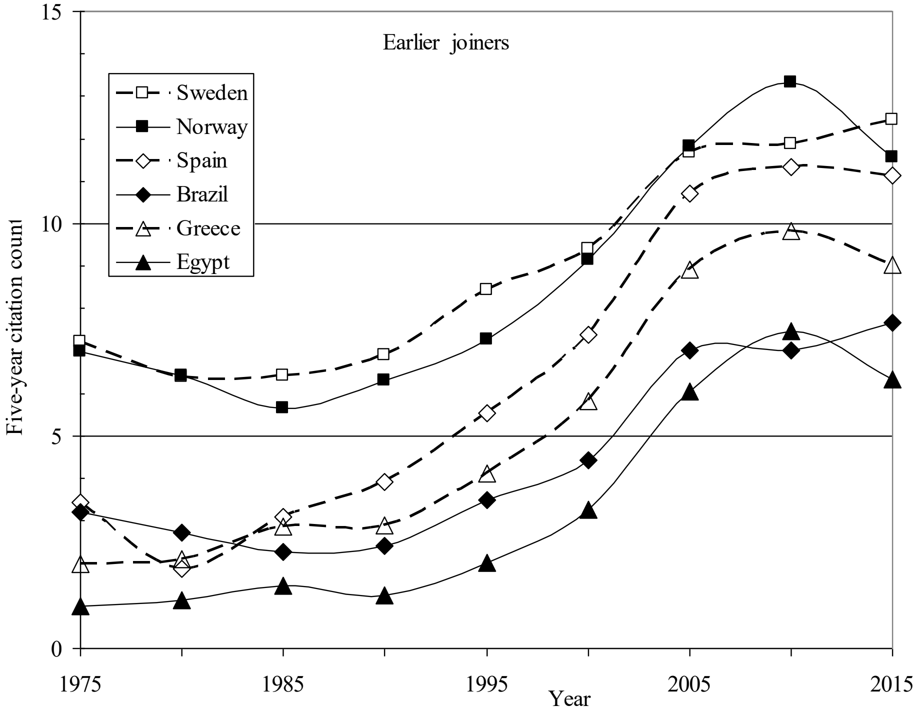


Figure 6. Five-year counts (arithmetic means) of Web of Science citations to domestic biomedical research papers of three new EC Member States who joined in 1981, 1986 and 1995, and their comparator non-EU MS countries.

Closer analysis (Table 2) shows that at ten years after each of the six new MS (five years only for Croatia) joined the EC or the EU, the ratio of additional papers in collaboration with other EU MS has improved for five of the six comparisons. The exception is Poland compared with Russia, where the ratio improved at five years after accession (2009) but declined at ten years (2014).

Citation scores for domestic papers

The five-year mean citation scores for the domestic papers for the six new EU MS and their comparator countries are shown graphically in Figures 6 and 7.

Discussion

This study has attempted to show that new Member States of the EU have gained advantages in their biomedical research activities through membership that were greater than similar changes that occurred with non-member countries. It is perhaps surprising that such advantages are even detectable because the amounts of financial support for research programmes, although measured in the billions of Euros, are really quite small in comparison with the total support provided by most national governments, private-non-profit (Knabe & McCarthy, 2012) and commercial sources. Thus Begum *et al.* (2018) showed that the support from the European Commission in 2009-13 for European cancer research only amounted to 2.4% of the overall total. In diabetes research, it was 2.7% (Begum *et al.*, 2017a).

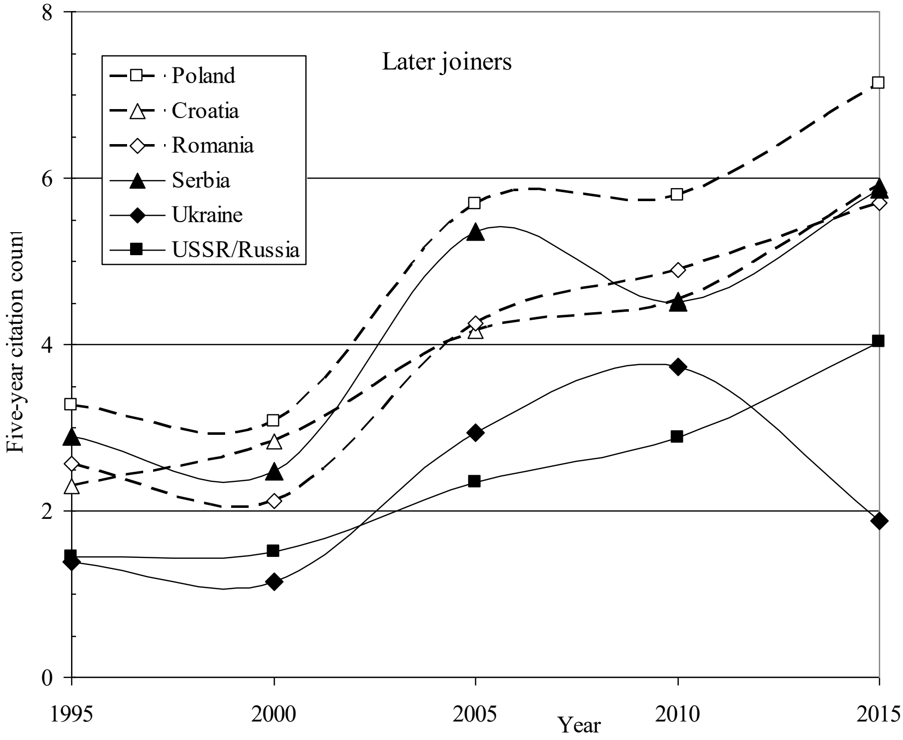


Figure 7. Five-year counts (arithmetic means) of Web of Science citations to domestic biomedical research papers of three new EC Member States who joined in 2004, 2007 and 2013, and their comparator non-EU MS countries.

Table 3. Table of five-year citation counts (ACI) for the 12 countries in the nearest year to their EU accession and 10 years later.

Country	*ISO*	*Joined*	*Near yr*	*ACI 0*	*+ 10 yrs*	*ACI +10*	*Change*	*Difference*
Greece	GR	1981	1980	2.11	1990	2.91	0.80	0.67
Egypt	EG			1.13		1.26	0.13	
Spain	ES	1986	1985	3.09	1995	5.53	2.44	1.23
Brazil	BR			2.29		3.5	1.21	
Sweden	SE	1995	1995	8.43	2005	11.67	3.24	-1.29
Norway	NO			7.29		11.82	4.53	
Poland	PL	2004	2005	5.69	2015	7.14	1.45	-0.24
Russia	RU			2.34		4.03	1.69	
Romania	RO	2007	2005	4.26	2015	5.71	1.45	2.51
Ukraine	UA			2.93		1.87	-1.06	
Croatia	HR	2013	2010	4.54	*2015*	5.91	1.37	0.02
Serbia	RS			4.51		5.86	1.35	

However, this is not the whole story. There are many other European activities that encourage the internationalisation of research, for example the several exchange programmes such as ERASMUS for students, and the Marie Curie programme for researchers. There is also the COST programme that brings researchers together for conferences and seminars. Closer links between EU MS are also encouraged by the European Regional Development Fund, which has improved the transport infrastructure (roads and railways), and the Commission's campaigns to reduce (since 2006) and abolish (since 2017) the trans-national roaming charges paid by users of mobile telephones. Some of the support also comes from the Cohesion Fund which benefits poorer MS. There has been a parallel effort to create a single market for aviation within Europe, which has increased competition, lowered fares, and removed the anti-competitive practices of national carriers found in, for example, Latin America. Safety has been improved by the Single European Sky (SES) programme. These moves have all made international collaboration in research much easier and cheaper.

The Single Market has also required public positions within the EU to be open to citizens of any MS. This has had the effect of improving competition and so driving up standards, including in research. For example, cancer researchers located in the UK in 2014-16 included almost 40% of people with non-British names (Begum et al., 2017b), and half of them were from other European countries. So this is one plausible means whereby the impact of domestic biomedical research may have increased more rapidly in EU MS than in the comparator countries.

There are two exceptions to this relative improvement in citation impact by EU MS relative to their comparator countries. Norway appears to have out-performed Sweden, and Russia to have out-performed Poland. In fact, Norway as a participant in the European Economic Area (EEA) Agreement (which also applies to Iceland and Liechtenstein) has for many years enjoyed the same free movement of people as EU MS. It is also particularly welcoming to foreigners, as measured by its 11th place in the Migrant Acceptance Index compiled by the Gallup polling organisation in 2016. It has also had an active strategy for international collaboration, and its wealth has facilitated this. The situation in Russia is rather different, because its scientific output and standing dropped very suddenly in 1990-95 as a result of the collapse of the Soviet Union. Its domestic biomedical research output declined to a nadir in 2000-07, and has been rescued by support from the European INTAS programme which started in 1993, and from George Soros in the USA. It may also have benefited more than

Poland from the addition to the WoS of the Emerging Sciences Citation Index (ESCI) in 2015, which had varying effects on different countries. Thus the ESCI papers from Poland in 2015-19 were only 11% of the total in the Science Citation Index – Extended plus the Social Sciences Citation Index, but in Russia they amounted to over 29%, and might have led to more citations, particularly from Russian authors.

In summary, the improvements in biomedical research collaboration between the new EU MS and the others, and in the citation impact of their domestic research, can reasonably be attributed to the activities of the European Commission over the last four decades. What we need to do next is to investigate how often research papers that involve two or more EU MS acknowledge financial support from the EU in its many different forms. This would be a major bibliometric study, not least because such support can be acknowledged in literally thousands of different formats.

As a preliminary exercise, we examined the acknowledgements on German papers in colorectal cancer research in the eight years, 2009-16. Ones written in collaboration with other EU MS were more likely to be in receipt of EU funding, see Table 4.

Table 4. Percentages of EU funding for German-authored papers on colorectal cancer, 2009-16, with different numbers of the nine leading other Member States who co-authored them. Mean five-year citation counts also shown for papers from 2009-15. ***Note: some of the papers with "0" other EU MS may have had co-authors from MS that were not analysed.***

Number of other EU MS	*Papers*	*With EU funding*	*%*	*Mean ACI*
0	653	22	3.4	26
1	347	45	13.0	33
2	99	17	17.2	32
3	65	13	20.0	96
4 to 6	84	14	16.7	83
7 or more	60	43	71.7	54

This table shows that papers from two or more EU MS are not only more likely to acknowledge EU funding, but also to be more frequently cited. Moreover, the presence of EU funding seems to attract a lot of additional support. For example, German colorectal research papers with EU funding attract an average of 16 other funders, but those without EU funding average only 2.8 funders in total, and 19% have no financial acknowledgements.

This study has some limitations. Time and resource constraints meant that the determination of ACI values was only performed for every fifth year, so that the results show some statistical scatter. We also limited the numbers of new MS studied to just six, and in particular omitted almost all of the cohort of MS that joined at the same time as Poland in 2004. The comparator countries were often not very like the MS, and their relationships with the EU varied greatly. They may also have varied in their relative commitments to biomedical research as a fraction of their GDP. We were also unable to measure the precision and recall of the biomedical filter as it was based on address terms, and these provide the means whereby subject-based filters are normally calibrated (Lewison, 1999).

Acknowledgments

The funding analysis of the German colorectal cancer research papers was carried out by Mursheda Begum, now of Queen Mary University of London. We are most grateful to Dr Charles Kessler, previously a scientist on the staff of the European Commission, for very helpful comments.

References

Begum, M Lewison, G Wright, JSF et al. (2016) European Non-Communicable Respiratory Disease Research, 2002-13: Bibliometric Study of Outputs and Funding. *PLoS ONE*, 11(4): e0154197

Begum, M Lewison, G Sommariva, S et al. (2017a) European diabetes research and its funding, 2002-2013. *Diabetic Medicine*, 34(10): 1354-1360

Begum,M Roe, P Webber, R & Lewison, G. (2017b) UK ethnic minority cancer researchers: their origins, destinations and sex. *Proceedings of 15th Conference of the International Society for Scientometrics and Informetrics*, Wuhan, China: 568-597

Begum, M Lewison, G Lawler, M & Sullivan, R. (2018) Mapping the European cancer research landscape: An evidence base for national and Pan-European research and funding. *European Journal of Cancer*, 100: 75-84

Begum, M Lewison, G Wölbert, E et al. (2020) Mental health disorders research in Europe, 2001-2018. *Evidence-Based Mental Health*, 23: 15-20

David, A (2016) The Participation of Austria and Hungary in the Framework Programmes for Research and Technological Development of the European Union. A Comparative Analysis. *Romanian Journal of European Affairs*, 16 (4): 48-67

de Nettancourt, D & Klepsch, A (1990) Research-and-Development Programs of the European-Communities in Biotechnology and Related Areas. *Food Biotechnology*, 4 (1): 625-634

Evangelatos, N Satyamoorthy, K Levidou et al. (2018) Multi-Omics Research Trends in Sepsis: A Bibliometric, Comparative Analysis Between the United States, the European Union 28 Member States, and China. *Omics-a Journal of Integrative Biology*, 22 (3): 190-197

Hefler, L Tempfer, C & Kainz, C (1999) Geography of Biomedical Publications in the European Union, 1990-98. *Lancet*, 353 (9167): 1856

Kalo, Z van den Akker, LHM Voko, Z et al. (2019) Is there a Fair Allocation of Healthcare Research Funds by the European Union? *PLoS ONE*, 14 (4

Knabe, A & McCarthy, M (2012) Civil Society Organisations and Public Health Research - Evidence from Eight European Union New Member States. *Central European Journal of Public Health*, 20 (4): 287-293

Lewison, G (1994) Publications from the European Community Biotechnology Action Program (BAP) - Multinationality, Acknowledgment of Support, and Citations. *Scientometrics*, 31 (2): 125-142

Lewison, G. (1999) The definition and calibration of biomedical subfields. *Scientometrics*, 46(3): 529-537

Lewison, G & Paraje, G. (2004) The classification of biomedical journals by research level. *Scientometrics*, 60(2): 145-157

Leydesdorff, L (2000) Is the European Union Becoming a Single Publication System? *Scientometrics*, 47 (2): 265-280

Leydesdorff, L Wagner, CS & Bornmann, L (2014) The European Union, China, and the United States in the Top-1% and Top-10% Layers of Most-Frequently Cited Publications: Competition and Collaborations. *Journal of Informetrics*, 8 (3): 606-617

Mela, GS & Cimmino, MA (1998) An Overview of Rheumatological Research in the European Union. *Annals of the Rheumatic Diseases*, 57 (11): 643-647

Pallari, E Soukup, T Kyriacou, A & Lewison, G. (2020) Assessing the European impact of alcohol misuse and illicit drug dependence research: clinical practice guidelines and evidence-base policy. *Evidence-Based Mental Health*, 23: 67-76

Parodi, S Parodi, A Lombardo, C et al. (1993) Cancer Research in the European-Community and Other Non-EC Countries. *Tumori*, 79 (1): 9-15

Ramos, JM Gutierrez, F Masia, M et al. (2004) Publication of European Union Research on Infectious Diseases (1991-2001): A Bibliometric Evaluation. *European Journal of Clinical Microbiology & Infectious Diseases*, 23 (3): 180-184

Sautter, J Olesen, OF Bray, J et al. (2011) European Union Vaccine Research-an Overview. *Vaccine*, 29 (39): 6723-6727

Soteriades, ES & Falagas, ME (2005) Comparison of Amount of Biomedical Research Originating from the European Union and the United States. *British Medical Journal*, 331 (7510): 192-194

Tikal, S (1993) European Communities Policy in the Sphere of Research-and-Development. *Ekonomicky Casopis*, 41 (11-12): 815-830

Weydert, M (2000) Research in the European Union: Continued Implementation of the 5th Framework Program. *Sea Technology*, 41 (1): 20-21

How are data paper abstracts constructed? Preliminary analysis of rhetorical moves in data paper abstracts from *Scientific Data* and *Data in Brief*

Kai Li[1] and Chenyue Jiao[2]

[1] *kai.li@ruc.edu.cn*
Renmin University of China, School of Information Resource Management, 15 Zhongguancun St., Beijing, China, 100080

[2] *cjiao4@illinois.edu*
University of Illinois Urbana-Champaign, School of Information Sciences, 501 E. Daniel St., Champaign, IL, United States, 61820

Abstract

The data paper is an emerging academic genre that responds to the rising importance of data in the scientific enterprise. It is a type of academic publication that focuses on the description of data objects. With a large number of data papers published in recent years, it is critical to understand their presence in the scholarly communication system. Our project aims to achieve this goal by investigating the rhetorical functions played by data papers. This research-in-progress paper reports preliminary results from this project. In this work, we expanded an established classification system of rhetorical moves in research article abstracts to make it applicable to data paper abstracts. We further applied this system to classify all sentences in 360 abstracts of data papers in two leading data journals, *Scientific Data* and *Data in Brief.* We identified four new rhetorical moves specific to data papers and examined how all rhetorical moves are distributed across the two journals. This work illustrates some important characteristics of data papers as a rhetorical device and informs future research directions towards a more comprehensive appreciation of data papers in the scientific system.

Introduction

Data have risen to be one of the most prominent epistemic objects in the scientific enterprise (Wynholds, 2011). The growing amount of data paves ways for new research questions and methods as well as large-scale scientific collaborations (Hey et al., 2009). All these changes, in return, require higher transparency of data in the scholarly ecosystem, as summarized into the FAIR Principles of research data stewardship, i.e., research data should be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016).

One approach to addressing the new needs for research data in the scholarly system is through the concept of *data publication*. Data publication refers to the pipeline through which data are transformed into discrete, well-documented, publishable, and citable objects (Parsons & Fox, 2013). An implementation of this concept that is gaining momentum is to publish data objects as academic papers, as shown in the emerging academic genre of *data papers*. A data paper is defined as a "scholarly publication of a searchable metadata document describing a particular online accessible dataset, or a group of datasets, published in accordance with the standard academic practices" (Chavan & Penev, 2011, p. 3).

Data papers are becoming more popular from the mid-2010s, partly supported by the establishment of journals dedicated to this genre, or *data journals*. Candela and colleagues' survey (2015) identified seven dedicated data journals and over a hundred academic journals accepting data papers along with research articles. In 2014, the journal *Scientific Data* was developed by Nature Publishing Group (Hrynaszkiewicz & Shintani, 2014) and *Data in Brief* by Elsevier (Thelwall, 2020), both later grew into the most important exclusively data journals in the market (Walters, 2020).

Given the short history of data papers, we are still far from a clear understanding of their presence in the scholarly communication system. One critical aspect of this knowledge is how these publications are constructed differently from research articles from a rhetorical perspective. In this regard, data papers offer a valuable site to observe the diversity of scientific rhetoric. On the one hand, data papers have distinct purposes from research articles: they are only supposed to describe the datasets per se, instead of offering information about the research design and results (Callaghan et al., 2012). On the other hand, data papers inevitably share many similarities with research articles, as both genres are produced in the same scholarly system and highly similar research contexts (Li et al., 2020). However, how these two genres are compared with each other has never been examined by empirical studies.

Our project aims to fill this gap by evaluating various aspects of rhetoric in data papers and investigate how they are connected to the roles played by data publications in scholarly communication. In the present work, we report our preliminary findings by identifying and classifying rhetorical moves in abstracts of a selected sample of data papers in two flagship data journals, *Scientific Data* and *Data in Brief*. The rhetorical move is commonly defined as a "discoursal or rhetorical unit that performs a coherent communicative function in a written or spoken discourse" (Swales, 2004, p. 268). Thus, this work will contribute to a deeper understanding of data publications from a socio-linguistic perspective. Specifically, the following two questions are pursued in this research-in-progress paper:

RQ1: What rhetorical moves are used in abstracts of data papers? This question aims to examine all rhetorical moves used in data paper abstracts and identify new moves that are specific to data-related contexts. We used the manual coding method and applied a modified classification system to identify all moves from abstracts of a selected sample of data papers.

RQ2: How are these moves distributed in the two data journals? This question strives to understand the cross-journal differences in the use of rhetorical moves. In this preliminary study, we only connected the differences to the journal policies concerning abstracts to draw preliminary explanations.

Method

In this research-in-progress paper, we only used data papers included in *Scientific Data* and *Data in Brief*, two domain-independent flagship data journals, from Scopus on November 15, 2020. While both journals were categorized as *exclusively data journals* by Stuart (2017), they may contain document types beyond data papers, especially comments and reviews. We removed all other document types and retrieved 7,712 data papers from these journals, with 6,335 from *Data in Brief* and 1,377 from *Scientific Data*.

These two journals were selected due to the following reasons. First, they are the two leading, domain-independent exclusively data journals based on the journal impact factor, the number of publications, and the presence in empirical studies (Stuart, 2017). Second, both journals were founded in 2014, which facilitates meaningful comparisons over time. Third, these journals have a few differences in the author guideline, which can shed light on the diversity of data papers. Based on these reasons, we believe our sample is able to reflect how data papers in a broad array of research domains are composed.

This preliminary study focuses on the journal policies of the paper abstract, which are summarized in Table 1. While rules on both journals are similarly worded, especially what should and should not be included in the paper abstract, one major difference is that *Scientific Data* directly states that no reference should be cited in the abstract.

We selected all data papers published between 2015 and 2020, as both journals were founded in 2014 and there are not enough papers to be analyzed in that year. To enable a more meaningful comparison between the journals, we used a stratified sampling approach where we take 30 publications for each journal in each year. There are finally 360 data papers in our sample for manual classification, to be described in the next section. We used the NLTK package of the Python language (Loper & Bird, 2002) to parse paper abstracts into sentences. A total of 2,182 sentences were parsed from all selected data papers.

Table 1. Abstracted-related policies from the two journals

Scientific Data[1]	*Data in Brief*[2]
- They should succinctly describe the study, the assay(s) performed, the resulting data and their reuse potential. - It should not make any claims regarding new scientific findings. - No references are allowed in this section.	- Concisely describes the data, its collection process, analysis and reuse potential. - Do not: provide conclusions, results, or mention the word 'study'.

Classification of rhetorical moves in abstracts

We classified all parsed sentences based on its rhetorical function(s) in the texts. For this purpose, we modified the classification scheme of rhetorical moves in research article abstracts proposed by Hyland (2000) that is composed of *Introduction*, *Purpose*, *Method*, *Product*, and *Conclusion*. Based on existing empirical evidence, these moves are not sufficient for data papers, due to the distinct functions of the latter genre (Li & Chen, 2018). As a result, we expanded this system using 50 randomly-selected abstracts (from the original 7,712 papers), where two coders independently reviewed them to identify any additions to this system. In the end, we added four new moves that are specific to data papers to Hyland's scheme. Our modified scheme is illustrated in Table 2, with the four added categories highlighted. Moreover, we changed *Product* into *Results* in our scheme, even though we retained its original definition.

Table 2. The new classification of rhetorical moves in data paper abstracts

Move	*Definition*
Introduction	Context of the papers
Purpose	Purpose or intention of the paper/research
Method	Research design, procedure, assumptions, approach of the study
Results	Main findings or results
Conclusion	Interpretations of the results beyond the scope of paper
Data description	**Description of the data object that is the topic of the paper**
Data uses	**How the data object is supposed to be used or its implications**
Data accessibility	**How to get access to the data object**
Related research article	**The research article to which the data object is connected.**

In this classification system, we specifically separated moves that are focused on the data object per se and those on the research behind the data object, as the data-research dichotomy is an important distinction between data papers and research articles. In the former group, four added categories specifically focus on any information about the data object being described in the data paper.

On the other hand, while some traditional rhetorical moves are supposed to be used in the data paper abstracts, especially *Introduction* and *Method*, moves like *Results* and *Conclusion* are clearly discouraged to be used based on the journal policies in Table 1.

Using the modified classification scheme, two coders independently classified all sentences. The intercoder reliability between the two coders is 0.706, indicating a good agreement (Landis & Koch, 1977). All differences between the coders were resolved before data were analyzed.

In our coding, we allowed the co-existence of multiple moves in the same sentence, given the complexity of human language. In our final result, we identified 77 sentences with two moves.

[1] https://www.nature.com/sdata/publish/submission-guidelines

[2] https://www.elsevier.com/journals/data-in-brief/2352-3409/guide-for-authors

We used a fractional counting method for these sentences in the next section, with a sentence being counted as 0.5 for each move its covers.

Results and discussion

Table 3 summarizes the counts of sentences and papers with the nine moves. The table shows that *Introduction*, *Method*, *Data description*, and *Data uses* are the most frequently used moves on both the sentence- and paper-levels. This group of moves is composed of both research- and data-oriented functions. A notable finding, contrasting to the journal policies, is that *Results* and *Conclusion* are often used in abstracts, especially that more than 50% of articles have at least one *Results* sentence. Following our previous work (Li et al., 2020), this finding, again, sheds doubt on the boundaries between research articles and data papers.

Table 3. Counts of sentences and papers with moves

Move	*# Sentences (fractional count; n = 2,182)*	*# Papers (n = 360)*	*Sentences per paper*
Introduction	514	217	2.37
Method	527	249	2.12
Data description	357	233	1.53
Data uses	227.5	188	1.21
Results	185.5	97	1.91
Purpose	149	139	1.07
Related research article	89.5	90	0.99
Conclusion	67.5	52	1.30
Data availability	61	67	0.91

It is also worth noting that when these moves are used in a paper, there is a large variance in terms of the number of sentences with the specific move. For example, *Introduction* and *Method* have over two sentences per abstract as compared to *Related research article* and *Data availability* with lower than one. The latter moves have fewer than one sentence per abstract because they can be co-used with other moves in the same sentence. One obvious explanation for such differences is the different amounts of details related to all rhetorical moves: research-oriented moves tend to be richer in details than their data-oriented counterparts.

A major interest of this paper is to investigate how the rhetorical moves are used differently across the two journals. Table 3 illustrates the number of papers with specific moves in the two journals. There are two stark differences between the journals. First and foremost, *Related research article* is only used in *Data in Brief*. This can be explained by the fact that nearly all such sentences are accompanied by an in-text citation, which is disallowed by *Scientific Data*. Second, both *Introduction* and *Data uses* are adopted much more heavily in *Scientific Data* than *Data in Brief*. For both moves, their different usages in these journals cannot be explained by policies concerning paper abstracts, but these may be connected to the journals' different paper structures and peer review criteria, which warrants a future research.

Moreover, we also examined how these moves are used in these journals over time. A few key findings emerge from Figure 1. First, most of the moves are used similarly in both journals and consistently used over time. Second, for the three moves that are used differently across the two journals, there seems to be a generally converging trend between the journals. For example, the use of *Related research article* has been decreasing in *Data in Brief* over time and the opposite trend can be observed for *Data uses*.

Table 3. Counts of papers with moves in the data journals (for each journal, n = 180)

Move	*Scientific Data*	*Data in Brief*
Introduction	141	76
Method	119	130
Data description	132	101
Data uses	135	53
Results	45	52
Purpose	62	77
Related research article	0	90
Conclusion	23	29
Data availability	36	31

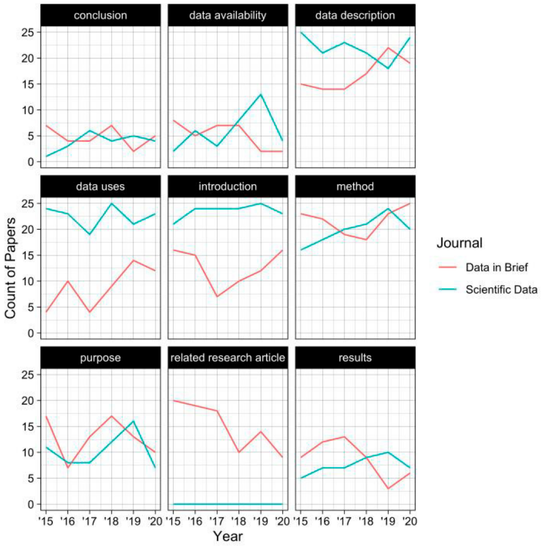


Figure 1. Number of papers with specific moves over time by journal

Conclusions

This research-in-progress paper is part of our larger research project aiming to investigate the rhetoric in data papers, so as to better locate this new academic genre in the scholarly communication system. One step towards this broad goal is to evaluate the differences between data papers and research articles in terms of used rhetorical moves. In this work, we present some preliminary findings concerning how rhetorical moves are used in data paper abstracts in two prominent data journals, *Scientific Data* and *Data in Brief*. We expanded an existing classification system of rhetorical moves in research article abstracts and identified four extra moves that are commonly used in the data-oriented contexts: *Data description*, *Data uses*, *Data accessibility*, and *Related research article*. We found that *Data description* is among the most frequently used rhetorical moves in our sample, along with *Introduction* and *Method*. Moreover, we also identified some notable differences in the use of rhetorical moves between the two journals, some of which can be explained by the differences in their journal policies about paper abstracts.

This work serves as a pointer to some important research directions to be pursued during the next steps of this project. First, it is important to connect the use of rhetorical moves to broader journal policy contexts and the domains in which data papers are produced. For example, some differences in our results may be explained by the different paper structures in these journals. Second, the combination and order of rhetorical moves in the abstract are better indicators of the story being told in academic publications. As a result, they will be studied in our future work to better understand key characteristics of data papers as a rhetorical device.

References

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A., Lowry, R., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & Wright, D. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *International Journal of Digital Curation*, *7*(1), 107–113. https://doi.org/10.2218/ijdc.v7i1.218

Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. *Journal of the Association for Information Science and Technology*, *66*(9), 1747–1762. https://doi.org/10.1002/asi.23358

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(15), 1.

Hey, T., Tansley, S., Tolle, K. M., & others. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Microsoft research Redmond, WA.

Hrynaszkiewicz, I., & Shintani, Y. (2014). Scientific data: An open access and open data publication to facilitate reproducible research. *J Inf Process Manag*, *57*, 629–640.

Hyland, K. (2000). Speaking as an insider: promotion and credibility in abstracts. *Disciplinary Discourses: Social Interactions in Academic Writing*, 63–84.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. http://www.ncbi.nlm.nih.gov/pubmed/843571

Li, K., & Chen, P. (2018). The narrative structure as a citation context in data papers: A preliminary analysis of Scientific Data. *Proceedings of the Association for Information Science and Technology*, *55*(1), 856–858.

Li, K., Greenberg, J., & Dunic, J. (2020). Data objects and documenting scientific processes: An analysis of data events in biodiversity data papers. *Journal of the Association for Information Science and Technology*, *71*(2), 172–182.

Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. *ArXiv Preprint Cs/0205028*.

Parsons, M. A., & Fox, P. A. (2013). Is data publication the right metaphor? *Data Science Journal*, *12*(0), WDS32--WDS46.

Stuart, D. (2017). Data bibliometrics: Metrics before norms. *Online Information Review*, *41*(3), 428–435. https://doi.org/10.1108/OIR-01-2017-0008

Swales, J. (2004). *Research genres: Explorations and applications*. Cambridge University Press.

Thelwall, M. (2020). Data in Brief: Can a mega-journal for data be useful? *Scientometrics*, *124*(1), 697–709.

Walters, W. H. (2020). Data journals: incentivizing data access and documentation within the scholarly communication system. *Insights*, *33*(1).

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.18

Wynholds, L. (2011). Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects. *International Journal of Digital Curation*, *6*(1), 214–225. https://doi.org/10.2218/ijdc.v6i1.183

Learning to represent for discipline

Yuan Lin[1], Zhang Xie[2], Haifeng Liu[3i*], Kun Ding[4], Kan Xu[5] and Shiyu Zhang[6]

[1,4,6] {zhlin,dingk,zhangshiyu}@dlut.edu.cn, [2] xzreal_dlut@163.com
Dalian University of Technology, WISELab, 2 Linggong Road, Ganjingzi District, Dalian, Liaoning (China)

[3] liuhaifeng@mail.dlut.edu.cn, [5] xukan@dlut.edu.cn
Dalian University of Technology, DUTIR, 2 Linggong Road, Ganjingzi District, Dalian, Liaoning (China)

Abstract

Interdisciplinary classification systems provide data sources for interdisciplinary research, and the representation learning can extract relevant subject characteristics at a deep level. This paper established four interdisciplinary classification systems and proposed two discipline representation model through knowledge representation learning and network representation. The results show that discipline groups can be merged into almost eight clusters, and discipline entity can be used to identify the compendex. The method of discipline expression not only helps to scientifically explore the inherent laws of disciplines, but also can quantitatively measure the relationship between disciplines.

Introduction

Discipline is the structure hierarchy of science, and a branch of scientific knowledge with specific research object. It could accelerate the comprehension and construction of science mapping to explore science inner construction and establish a comprehensive and systematic scientific structure system for researcher. Measuring interdisciplinarity can reveal characteristics of knowledge flow. As scientific research extends beyond a single discipline, many scientists have realized the potential of interdisciplinary research.

Bibliometrics are commonly used to measure the interdisciplinarity of research by citation analysis, which can show present, past and future activity in science. Through citation analysis, some researchers explored and constructed many categories classification systems, such as Journal Citation Reports (JCR), Essential Science Indicator (ESI), Ei Compendex (EI) and Scimago Journal & Country Rank (SJR). Thousands of journals were divided into different research areas with specific rules in four category systems.

Academic journals are important carriers of scientific activities and important academic communication media for scientific development. Journal classification of research unit in scientific knowledge represent discipline, which could promote scientific knowledge management. Scientific and technical workers usually publish the new discoveries, new laws, new theories and new methods in academic journals. Cluster analysis based on citation relations with journals can explore the relationship and structure between related disciplines, analyse the structure of interaction and the trend of penetration and derivation, so as to reveal the dynamic structure and development law of science.

In the past decades, many researches explored the internal rules of the discipline with different methods. Leydesdorff used citation analysis, multivariate analysis to classify journals , based on citation data in the database, and showed the change trajectory of discipline structure corresponding to different journal groups, revealed the dynamic structural characteristics of science. Porter proposed that integration is reflected in the diversity of the cited subject's categories as a compelling interdisciplinary research metric. Porter explored differences between similar or dissimilar interdisciplinarity by using the number of disciplines cited. Wagner gave a literature review of quantitatively measure of interdisciplinarity and pointed out that bibliometrics, such as co-authorship, co-inventors, collaborations, references, citations, and co-citations, were used to measure interdisciplinarity. Zhang considered the similarity between discipline areas through the diversity of references to research interdisciplinarity of journal.

Solomon explored multidisciplinary journal interdisciplinary between Nature and Science. The above researches had usually been used to measure the interdisciplinarity in journals.

As early as the last century, the founder of scientometrics, Price believed that citation data from journals could provide useful information for scientific structure researches and scientific classification systems, and journals were a very suitable research unit for the research of scientific structure from the perspective of citation analysis. It inspired the idea of using journals as units of science to study interdisciplinary research.

With the development of information technology, it is possible to analyse data and mine information with network graph. The representation learning method is used to mine multiple information entities to explore potential information. Representation learning was used to map symbolized data to a low-dimensional space to obtain information. Representation learning expresses the wide application of learning methods in the field of data analysis greatly and assists the analysis of multi-source information. Along with the vigorous development of deep learning technology, representation learning methods have developed rapidly in recent years, meanwhile, such methods have greatly facilitated the analysis of multi-source information.

The representation learning method combines multiple information entities to mine potential information. Representation learning is divided into knowledge representation learning and network representation learning.

Knowledge representation learning maps entities and relationships to low-dimensional dense vector space of the same dimension based on complex semantic associations between entities and relationships which are calculated with the spatial distance. Typical methods include Word2vec proposed by Mikolov, which used low-dimensional vectors to learn and represent semantic information. Knowledge representation learning was also used widely in the mining of text semantic information to learn the relevant information between text information effectively. Bordes proposed to embed the materialized relation into the TransE model in the low-dimensional vector space.

Network representation learning represents the semantic information of complex network nodes as dense low-dimensional real value vectors, which can calculate the semantic connection and solve the problem of semantic association extraction and computational complexity under sparse data effectively between nodes in the network , also can improve knowledge acquisition and reasoning performance significantly. Typical network representation learning took nodes as words by forming a network representation learning method based on node location information represented by deepwalk and node2vec. The above methods obtain the sequence data of the nodes in the complex network by random walk. These node sequences can be analogized to the sentences in word2vec, and the nodes are analogous to the words in the sentence, and then the node sequence data is input into word2vec model. The representation vector can obtain network information of each node. For example, Zhang used network representation learning to represent literature information and predict scientific research cooperation. By analysing the author's cooperative relationship and clustering the author according to the research topic, the relative authors can be minded with same research topic, which revealed research hotspot and accelerated scientific research cooperation.

In interdisciplinary research, journals and disciplines are the basic components that together form a discipline-periodical relationship network. There are similar research directions in different disciplines. The deep features of the disciplines can be quantitatively expressed through network representation learning to complete internal information mining.

Methodology

Analysis of four interdisciplinary classification systems

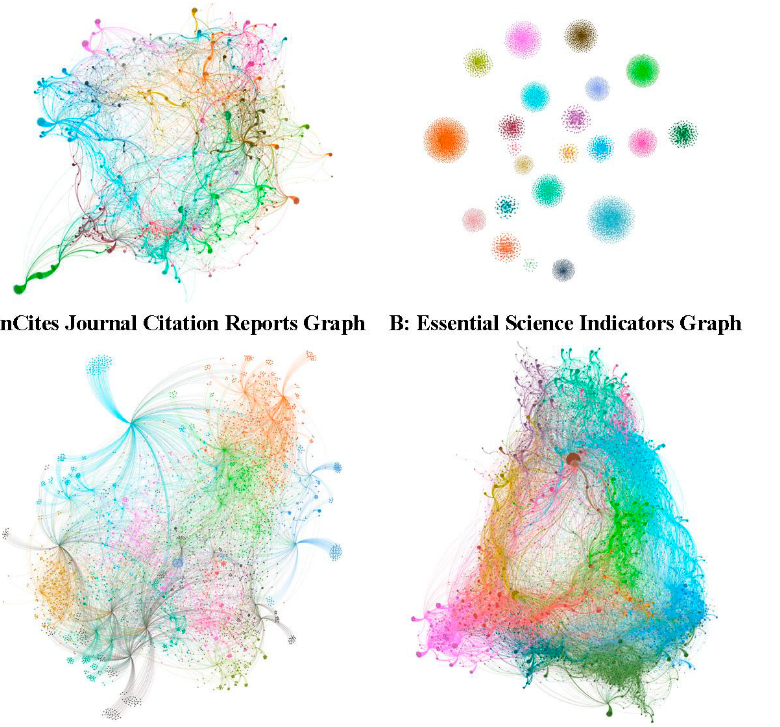
The current international interdisciplinary classification system includes the following four categories: Clarivate Analysis's Incites Journal Citation Report (JCR), Clarivate Analysis's Essential Science Indicators, Elsevier's Engineering Index (EI), and SCImago's Scimago Journal & Country Rank (SJR).

The discipline classification data of JCR is obtained by accessing the journal Citation of the Web of Science (JCR), the discipline classification data of ESI is obtained by accessing the module of Essential Science indicators of Web of Science (ESI), and the discipline classification data of SCImago Journal & Country Rank (SJR) and Ei Compendex (EI)is obtained by downloading their published journal list of discipline classification systems in the relevant home page.

JCR allows users to evaluate and compare journals using citation data from approximately 12,000 academic, technical journals and conference papers from more than 3,300 publishers in more than 60 countries and territories. Journal citation reports are the only source of journal citation data, covering almost all disciplines in science, technology and social sciences. ESI is an important analytical tool for measuring scientific research performance and tracking scientific development trends based on SCI and SSCI. As one of the important indicators for the evaluation of first-class disciplines, ESI highly cited papers are the top 1% papers in citation frequency of each discipline. Their quantitative characteristics can reflect the level of discipline development macroscopically and evaluate the discipline competitiveness of institutions. Its content characteristic may manifest the discipline research hot spot and the frontier, guides the discipline development direction. EI is an engineering research database created in 1884, which providing the most comprehensive searchable engineering literature and patent information index. Designed by the engineering community, this comprehensive platform provides peer-reviewed, in-depth index and accurate engineering research content for academic institutions, enterprises and government agencies. SJR is an open portal website that includes journals and national scientific indicators published in accordance with the Scopus database. These indicators can be used to assess and analyse the scientific field. Journals can be compared or analysed separately. Citation data is derived from 34,100 publications from more than 5,000 publishers and national performance indicators from 239 countries worldwide. According to the classification systems of JCR, ESI, EI and SJR, journals will correspond to one or more disciplines.

On the basis that the current four classification systems are standardized and reasonable, due to the existence of shared journals, explorations of crossover and integration can be carried out. The current work is to explore the integration of classification systems. Among different classification systems, we construct a discipline network through discipline-journal-weight tripartite. The edge of network is defined as: In the JCR classification system, if the A journal is included in the discipline C, the weight is 1; if other classification systems exist same discipline C, and the A journal is also divided into the discipline C, the weight is +1. According to the rules of the subordinate network, this paper constructs a discipline-journal map of four classification systems. Four discipline relationship graphs are shown in Figure 1.

In the Figure 1, different colors represent different discipline clusters, and node size indicates the importance of disciplines in the cluster. In the JCR classification system, different journals are divided into one or more discipline. According to the same journals included relation, a bridge is generated, and a complex interdisciplinary association is formed between different disciplines. Through the community discovery algorithm, the JCR classification system contains 8 major discipline clusters. In the ESI classification system, journals are divided into single discipline, so the discipline relationship graph is shown in the Figure 1.B forming 22 discipline clusters. In the EI classification system, 30 discipline clusters are identified and generated. A large number of journals belong to the only discipline of EI classification system, and discipline clusters are closely related to each other through cross journals. More than 28,000 journals are included in the SJR classification system, and the volume of journals is much larger than other classification systems. The interdisciplinary phenomenon is obviously observed in four classification graphs.



A: InCites Journal Citation Reports Graph **B: Essential Science Indicators Graph**

C: Engineering Compendex Graph **D: Scimago Journal & Country Rank Graph**

Figure 1. Different categories classification graph

In order to observe the current situation of scientific development comprehensively, the scientific graph of the merge multi-source classification system is constructed by merging the four categories classification graph. The comprehensive graph is shown in Figure 2.

The Figure 2 integrates four classification systems, where the community discovery algorithm is adopted to identify 18 disciplines clusters. Different colors represent different disciplines, and the node size indicates the node importance. In figure 2, Medicine, Social Science, Engineering and Mathematics occupy the dominance in the multitudinous discipline clusters, forming the intricate relationship among discipline clusters, but closely related disciplines can still be distinguished, and disciplines still have significant difference. We need to calculate the relative degree between the different disciplines quantitatively and excavate inherent law of science urgently.

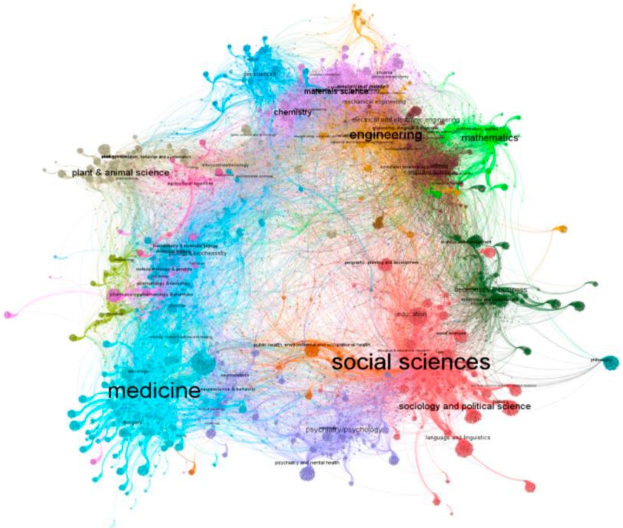


Figure 2. Comprehensive graph of four categories classification

Single-system discipline representation model

The single-system discipline representation model uses a discipline Vector Space Model to quantitatively analysis interdisciplinarity. The motivation of compendex included in a discipline is important to the development of discipline, regardless whether the journal is the core journal for other disciplines or not. All the core journals included in the same discipline have equivalent contribution to the discipline, and they have the same property for the discipline. Based on the assumption, we proposed the discipline vector space model to calculate the similarity between disciplines.

We constructed a 234-dimensional Vector Space Model to represent single-system discipline within the JCR interdisciplinary classification system, used cosine similarity to calculate the degree of interdisciplinarity. The vector of discipline is constructed by the number of multi-assigned journals among different disciplines. We used C discipline as a bridge to measure the degree of interdisciplinarity between A and B discipline. The number of Multi-assigned journals percentage between A and C as a dimension, and the number of Multi-assigned journals percentage between B and C as another dimension, then the relevance between A discipline and B discipline is calculated using these two dimensions as input for discipline vector space model. The construction of the vector is shown in the Table 1.

Table 1. Vector of The representation of Single-system discipline

Table	*ECONOMICS*	*MATHEMATICS APPLIED*	…	*Management*	*Law*
ECONOMICS	1	0	…	0.06	0.06
Mathematics	0	0.44	…	0	0.01

Multisource-system discipline representation model

Journals are classified to different disciplines according to the different discipline classifications system. The total number of disciplines is stable relatively, and the total number of journals is also stable relatively. The path of knowledge dissemination of science and the inherent law of science are embedded in the network of disciplines and journals. We used graph representation learning to solve the comprehensive graph embedding, comprehensive discipline embedding could excavate the potential law within the scientific system, calculate the relationship quantitatively between different disciplines, recognize the discipline compendex. The graph representation method is shown in Figure 3.

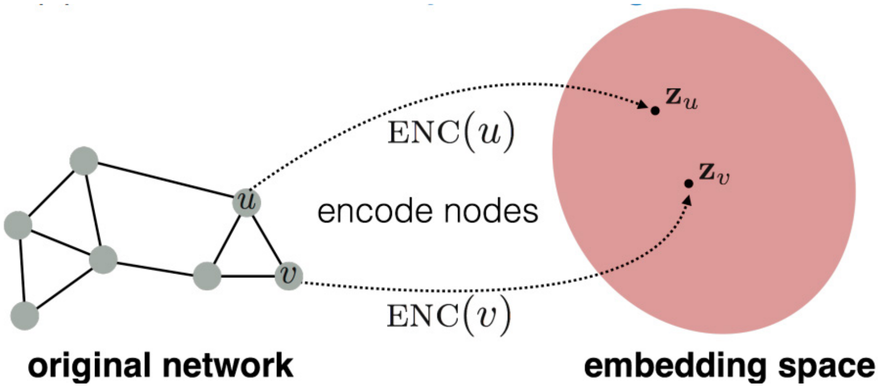


Figure 3. Network representation learning model

The network representation learning gold is to encode nodes so that the similarity in the embedded space is similar to the similarity in the original network. Two key components are encoder and similarity function. The encoder maps each node to a low-dimension vector.

$$\mathrm{ENC}(v) = Z_v \tag{1}$$

Where v is node in the input graph and Z_v is d-dimension embedding.

The similarity function is defined as follows:

$$\text{similarity } (u, v) \approx Z_v^T Z_u \tag{2}$$

Similarity is the edge weight between u and v in the original network. In order to train graph embedding, Loss function was used to calculate information loss, the loss function definition is as following :

$$\text{Loss} = \sum_{(u,v)\in VX\,V} \left|\left|Z_u^T Z_v - A_{u,v}\right|\right|^2 \tag{3}$$

Where the loss is what we want to minimize, $Z_u^T Z_v$ is embedding similarity and $A_{u,v}$ is adjacency matrix for the graph.

Graph representation learning in this paper adopts the node2vec(Grover and Leskovec 2016) model which used the Skip-gram method to extract representation for networks, and strike a balance between depth priority and breadth priority by using Random Walk. In the node2vec, definite the $A_{u,v}$ as log [exp $(Z_u^T Z_v)/\sum_{n\in V}$ exp $(Z_u^T Z_n)$]. By using the Node2vec model, each node in the final network is represented as a dense vector of 128 dimensions, which contain the importance, impact and correlation of disciplines and journals, the discipline representation vector can be used as represent discipline. Discipline represent vector used in the task of discipline correlation strength measurement, cross-discipline prediction and so on. Table 2 shows the example of multisource-system discipline representation vector.

Table 2. Example of multisource-system discipline representation vector

Table	*Dimension 1*	*Dimension 2*	…	*Dimension 128*
Management(discipline)	-0.7661	-0.4544	…	-0.7481
Scientometrics(journal)	-0.5728	-0.1878	…	0.9431

After the network representation learning model used on the information of disciplines and journals, each node in the network will be represented by the unique 128-dimensional representation vector. The further analysis of the representation vectors of disciplines and journals can excavate the rich implicit information in the network.

Experiment

Discipline group identification

Through the single-system discipline representation, we constructed the JCR discipline representation embedding, and multisource-system discipline representation embedding was constructed in the comprehensive discipline network. The above two representations are used to obtain a comprehensive discipline representation vector, which containing single-system discipline representation and the multisource-system discipline representation information through vector merge. Figure 4 shows the process of specific discipline representation vector:
The discipline representation vector is composed of single-system discipline representation vector (comes from the discipline vector space model of knowledge representation learning) and the multi-system discipline representation vector (comes from the network representation learning). The single-system discipline representation vector equal to 234-dimensional, the multi-system discipline representation vector equal to 128-dimensional. The comprehensive discipline representation vector is a linear connection of the preceding two vectors, which equal to 362 dimensions.

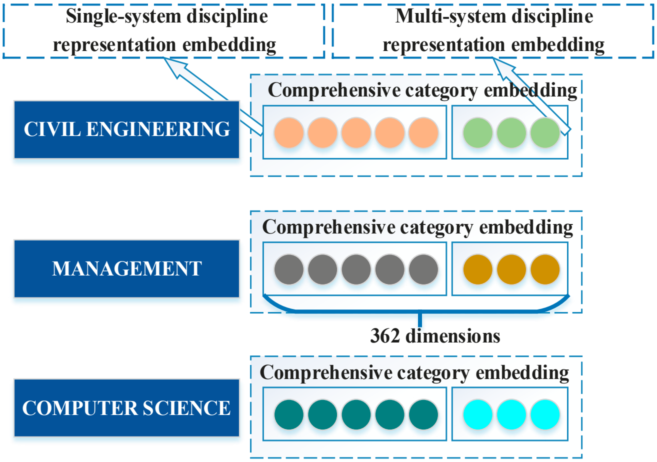

Single-system discipline representation embedding
Multi-system discipline representation embedding
CIVIL ENGINEERING
Comprehensive category embedding
MANAGEMENT
Comprehensive category embedding
362 dimensions
COMPUTER SCIENCE
Comprehensive category embedding

Figure 4. The representation vectors merge of discipline

Through network representation learning, a variety of information in the discipline network is considered and integrated comprehensively into the vector, each information representation entity is mapped into the space of the same dimension, and the distance between all the information representation entity is calculated by using the cosine similarity (the closer the distance, the higher the similarity between the entities). The similarity calculation method is shown as formula (4).

$$\text{Cosine}\,(\text{x},\text{y}) = \frac{X\cdot Y}{||X||||Y||} = \frac{\sum_{j=1}^{t} c_{xj}c_{yj}}{\sqrt{\sum_{j=1}^{t}(c_{xj})^2\sum_{j=1}^{t}(c_{yj})^2}} \tag{4}$$

x,y are two disciplines, X , Y are discipline comprehensive vectors, c_{xj} is the number of representation vector between x discipline and j discipline, t is the number of vector dimensions , which is equal to 362.The result is a positive number which is in the range [0,1] , and the result is 1 when the two vectors are exactly equal. The larger the similarity between the vectors is, the higher the value of cosine method will be. The closer the research direction between disciplines, the greater their similarity, vice versa.

According to the calculation method of the similarity degree of the discipline representation vector, this paper calculated all of similarities between disciplines in JCR and constructed a discipline similarity matrix of 234 times 234. In order to analyse the correlation intensity between disciplines in a balanced manner, this paper selected disciplines with interdisciplinary similarity greater than 0.5, and constructed a cross-discipline network. The weight of the edge is the value of similarity.

Meanwhile, we used Gephi visualization software to make a data perspective on the cross-discipline graph of JCR and used the complex network community detection algorithm to identify the discipline clusters. Different colours represent different discipline clusters. Node size indicates the importance of disciplines in the cluster.

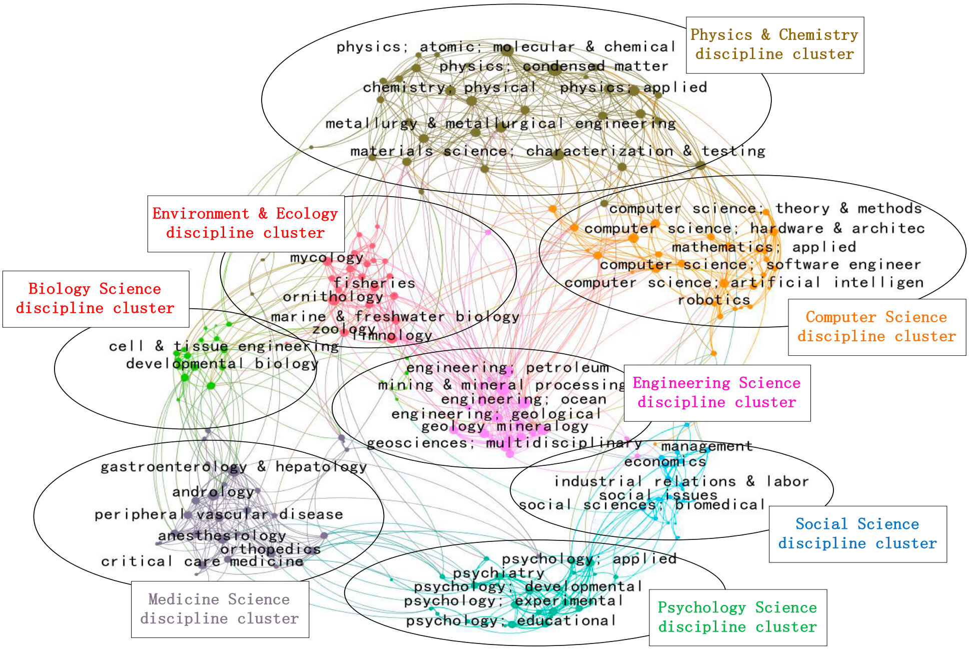

Physics & Chemistry discipline cluster
physics; atomic; molecular & chemical
physics; condensed matter
chemistry; physical
physics; applied
metallurgy & metallurgical engineering
materials science; characterization & testing
Environment & Ecology discipline cluster
mycology
fisheries
ornithology
marine & freshwater biology
zoology
limnology
Biology Science discipline cluster
cell & tissue engineering
developmental biology
computer science; theory & methods
computer science; hardware & architec
mathematics; applied
computer science; software engineer
computer science; artificial intelligen
robotics
Computer Science discipline cluster
engineering; petroleum
mining & mineral processing
engineering; ocean
engineering; geological
geology mineralogy
geosciences; multidisciplinary
Engineering Science discipline cluster
management
economics
industrial relations & labor
social issues
social sciences; biomedical
Social Science discipline cluster
gastroenterology & hepatology
andrology
peripheral vascular disease
anesthesiology
orthopedics
critical care medicine
Medicine Science discipline cluster
psychology; applied
psychiatry
psychology; developmental
psychology; experimental
psychology; educational
Psychology Science discipline cluster

Figure 5. The interdisciplinary graph of JCR classification

Observing the Figure 5, the current mainstream of the subject will be merged into 8 disciplines clusters, the eight-discipline cluster are Physics & Chemistry, Environment & Ecology, Computer Science, Engineering Science, Biology Science, Social Science, Medicine Science, and Psychology Science. By using the discipline representation vector, information about journals and disciplines were integrated into the discipline representation vector. We enlarged the scale of the discipline and journal network to the multisource discipline classification systems, integrated the information transmission of journals in the discipline-related network through multi-system discipline classification, and then the disciplines were used as a bridge to explore the representation methods of the disciplines. Finally, the single-system discipline representation vector was integrated with the multi-system discipline representation vector to obtain comprehensive discipline representation vector in Web of Science. By observing the Figure 5, the comprehensive discipline representation vector can calculate the intrinsic correlation information between disciplines more accurately, and community detection algorithm can make a perspective on the discipline-related network constructed by the representation vector to analyse the present situation of the integration of the disciplines. Today's disciplines have formed about 8 broad categories, different discipline clusters only through a few bridge nodes to generate information transmission. The ability to transmit information is relatively weak between disciplines that are far apart, hence information needs to be transmitted through multiple disciplines.

Discipline group identification

Discipline consists of several journals, and the composition system of the journals is the foundation discipline development. Measuring the degree of cross-correlation between journal and disciplines is important for discipline research. We constructed a discipline entity, mining the similarity between this entity and various journals, thereby discovering compendex.

Taking science of science as an example, science of science is a research field formed by the integration of natural sciences, social sciences, and humanities, and it is a comprehensive discipline to explore the coordinated development of science, technology, economy and society. In order to analysis science of science research direction accurately, through the four core journals (Scientometric, Journal of Informetrics, INFORMATION PROCESSING & MANAGEMENT, Journal of the Association for Information Science and Technology) as science of science source journals, construct science_of_science discipline and add into discipline representation network system, calculate the representation vector of science_of_science . We used the cosine similarity to calculate similarity between discipline representation vectors and journals representation vectors. The calculation method is similar to formula (1). Specially, 128 dimensions network representation vector were used to recognize science_of_science compendex. By calculating the similarity between the representation vector of science of science and all journals, most relevant journals about science of science are shown in Table 3. For ease of observation and analysis, the science_of_science compendex is represented in Figure 6.

In addition to the four original journals, many journals still have strongly relation to science of science. Science of Science top 10 relevance journals as shown in Table 3, through manual examination, other relative journals are also the mainstream journals in which science of science researchers submit research results.

The closer nodes to science of science in Figure 6, the greater the degree of similarity will be. It can be observed that original four journals (Scientometrics, Journal of Informetrics, INFORMATION PROCESSING & MANAGEMENT, Journal of the Association for Information Science and Technology) still have stronger correlation. According to the discipline and journal representation vector, we set similarity threshold as 0.89. Using network representation vector to recognize compendex in discipline could make up empiricism deviation problem.

Table 3. Science of Science compendex (TOP10)

Rank	*Discipline*	*Journal*	*Similarity*
1	science_of_science	aslib journal of information management	0.905
2	science_of_science	information development	0.902
3	science_of_science	econtent	0.899
4	science_of_science	interlending & document supply	0.899
5	science_of_science	journal of documentation	0.898
6	science_of_science	australian academic & research libraries	0.898
7	science_of_science	african journal of library archives and information science	0.897
8	science_of_science	knowledge organization	0.894
9	science_of_science	electronic library	0.889
10	science_of_science	library and information science	0.889

In this way, for any discipline or research field, it is possible to dig out the compendex in discipline and their correlation strengths in the discipline field by constructing the original discipline subordinate network by using the discipline representation vector model. In the process of identifying the compendex of the discipline, it can assist the scientific research scholars to further identify discipline hotspots and excavate discipline cooperation.

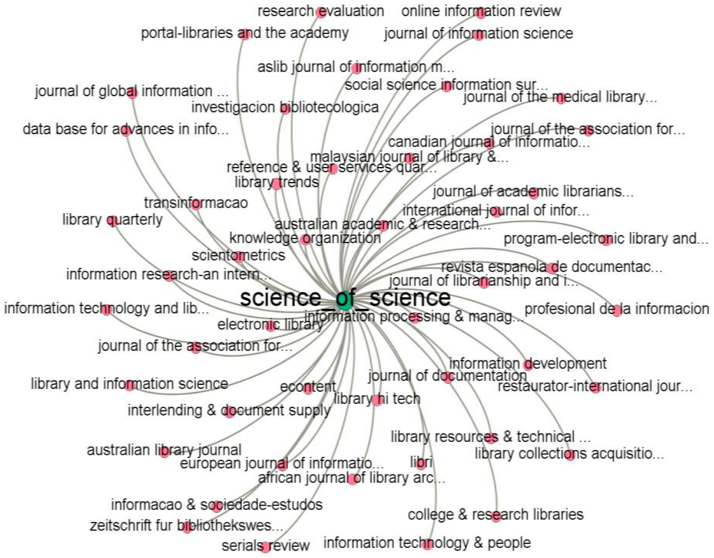

research evaluation
online information review
portal-libraries and the academy
journal of information science
aslib journal of information m...
social science information sur...
journal of global information ...
investigacion bibliotecologica
journal of the medical library...
data base for advances in info...
journal of the association for...
canadian journal of informatio...
malaysian journal of library &...
reference & user services quar...
library trends
journal of academic librarians...
transinformacao
international journal of infor...
library quarterly
australian academic & research...
knowledge organization
program-electronic library and...
scientometrics
revista espanola de documentac...
information research-an intern...
journal of librarianship and i...
science_of_science
information technology and lib...
profesional de la informacion
electronic library
information processing & manag...
journal of the association for...
information development
journal of documentation
library and information science
restaurator-international jour...
econtent
library hi tech
interlending & document supply
library resources & technical ...
library collections acquisitio...
australian library journal
european journal of informatio...
libri
african journal of library arc...
informacao & sociedade-estudos
college & research libraries
zeitschrift fur bibliothekswes...
serials review
information technology & people

Figure 6. Science of Science compendex graph

Conclusion

The construction of disciplines is inseparable from the establishment of a discipline classification system. Rigorous discipline division criteria and clear journal division standards are the basic support for discipline construction. The four interdisciplinary classification systems provide the possibility of integration due to the existence of interdisciplinary journals. This paper proposed two discipline representation models through four interdisciplinary classification systems, analysed the internal domain division of disciplines, constructed two types of discipline representation network. Discipline vector space model of knowledge representation learning model was used to obtain single-system discipline representation vector, network representation learning model was used to obtain comprehensive representation vectors of all disciplines and journals. The two models are dedicated to solving the problems of identification of interdisciplinary groups in the discipline measurement, quantitative calculation & recognition of discipline compendex.

In this process, we used the novel network representation learning model in the field of deep learning for discipline modelling. The applications of the discipline representation learning model proposed are also helpful for issues such as pre-evaluation of discipline development, discipline planning, and prediction of scientific evolution direction. In our work, the representation learning model proposed is not a replacement for traditional bibliometrics, but a method innovation that inherits the essence of classical bibliometrics. It is a preliminary exploratory study in the field of interdisciplinary classification systems, which is to provide a methodology for large-scale data mining when the explosive information growth. In the future work, we would incorporate more information to quantify the discipline representation， explore more effective methods to represent disciplines, integrate the existing discipline classification system for relevant analysis more closely , and summarize more enlightenment brought by the interdisciplinary.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 61976036.

References

Bengio, Yoshua, Aaron Courville & Pascal Vincent. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828.

Bordes A, Usunier N & Garcia-Duran A. (2013). Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2787-2795.

Cao, S., Lu, W., & Xu, Q. (2015). Grarep: Learning graph representations with global structural information, *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 891-900).

Frank, Roberta, Stanley Bailis, Julie Thompson Klein & Raymond Miller. (1988). Interdisciplinary: The First Half Century. *Issues in Interdisciplinary Studies*.

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks, *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 855-864).

Leydesdorff, L., & Cozzens, S. (1993). The delineation of specialties in terms of journals using the dynamic journal set of the SCI. *Scientometrics*, 26(1), 135-156.

Mikolov, Tomas, Kai Chen, Greg Corrado, & Jeffrey Dean. (2013). “Efficient Estimation of Word Representations in Vector Space.” *ArXiv Preprint ArXiv*:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 3111-3119.

Perozzi, Bryan, Rami Al-Rfou & Steven Skiena. (2014). “Deepwalk: Online Learning of Social Representations.”, *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 701–710).

Porter, A., Cohen, A., David Roessner, J., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, 72(1), 117-147.

Porter, Alan L. & Ismael Rafols. (2009). “Is Science Becoming More Interdisciplinary? Measuring and Mapping Six Research Fields over Time.” *Scientometrics*, 81(3), 719–45.

Price & Derek J. De Solla. (1965). “Networks of Scientific Papers.” *Science*, 149(3683), 510-515.

Solomon, G. E., Carley, S., & Porter, A. L. (2016). How multidisciplinary are the multidisciplinary journals science and nature?. *PLoS One*, 11(4), e0152637.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., ... & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of informetrics*, 5(1), 14-26.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology*, 67(5), 1257-1265.

[i*] Yuan Lin, Zhang Xie and Haifeng Liu contributed equally to this work.

Research on the Persona of Chinese Distinguished Young Scholars in Computer Science Area

Yuan Lin[1], Xinhang Zhao[2], Jiaojiao Feng[3i] and Kun Ding[4]

[1,4] *{zhlin,dingk}@dlut.edu.cn, [2]xinhang@mail.dlut.edu.cn, [3]joejoesky1993@163.com*
Dalian University of Technology, WISELab, 2 Linggong Road, Ganjingzi District, Dalian, Liaoning (China)

Abstract

Outstanding young scholars play an important role in national development, and thus their cultivation and funding are thriving research topics in academia. This paper studies the common characteristics of the Chinese scholars in the computer science area when they got funded by *the National Science Fund for Distinguished Young Scholars* project (i.e. CCSDY scholars) and creates a persona based on these characteristics. The persona of the CCSDY scholars in this paper includes personal attributes, educational attributes and academic attributes. The personal attributes and educational attributes are extracted from the data of their curricula vitae. The academic attributes are obtained by two processes: conducting a bibliometric analysis of their high-quality papers, and using the three-stage time series topic analysis model to get the most common research topics of these scholars in different time periods. This study accomplished a persona for the CCSDY scholars, reveals the crucial elements in their development and contributes to the guidance for countries to cultivate and fund scholars.

Introduction

As Chambers et al. (1998) argue: “Better talent is worth fighting for”. It is fundamental to cultivate outstanding academic scholars because their contributions are crucial to the development of countries. Therefore, the Chinese government has carried out various projects that facilitate young scholar cultivation and encourage overseas scholars to work in China (Zhu, 2019). It has been found that there is a positive correlation between paper usage (e.g. citation) and funding, which also shows the benefits of funds to the scholars from an academic perspective (Zhao et al., 2018). The funded projects financially support young scholars’ research, and in return, their research achievements can help China make progress in science and technology improvements (Yu et al., 2020). With the rapid growth of artificial intelligence, the computer science area is currently of great interest and attracts more and more scholars devoting themselves to it. And the Chinese scholars in computer science area also made notable contributions to the development of artificial intelligence (Lu et al., 2018).

Creating personas of target users is a widely-adopted method for building a concrete user representation (Pruitt & Grudi, 2003). Although persona is fictitious, it comes from the data of real users and can effectively embody users with demographic and biographical characteristics under modelling (Junior & Filgueiras, 2005). Persona is a useful approach to understand target groups and is mostly used in the user interface (Johansson & Messeter, 2005). In the age of big data, personas can be created with rich and easily accessible data (Guo & Ma, 2018). However, there remains a paucity of studies using personas to describe and model scholars so far.

This research gap is filled in this paper by combining CV analysis method to conduct a multi-dimensional analysis and accomplishing a persona of the scholars when they won *the National Science Fund for Distinguished Young Scholars* project (NSFC, 2018) in computer science area[ii] (i.e. the CCSDY scholars). To derive the research topics and presents their evolution, this paper proposes a three-stage time series topic analysis model (TTSTA model) to extract and analyse the research topics of the CCSDY scholars in three time periods: early, mid-term and recent. Recommendations for practice and policy in scholar cultivation and funding are also set out. The remaining part of this paper proceeds as follows: structure of scholar persona, data collection and processing method, multi-dimensional attribute analysis, result and conclusion.

Scholar persona

The researches on scholar characteristics mainly focus on three aspects: personal, educational and academic attributes. First, within personal attributes, the relationship between the scholars' age and research output has always been of interest (Frosch, 2011). Gender is also a well-studied element, and there have been reflections on the phenomenon of gender imbalance in academia (Lin, 2017). Second, regarding educational attributes, previous research has shown that the guidance of mentors plays a significant role in talent development (Margot & Kettler, 2019). Through the education experience analysis, many meaningful common characteristics were discovered, such as research-embraced culture, a dedicated mentor and partnerships (Archer, 2007). Finally, academic attributes is a growing research area, and the number of papers is the most frequent researched factor (Shen et al., 2016). As a step moved forward, researchers have paid more attention to the research interest of scholars. The topic mining algorithms include the LDA algorithm and the HDP algorithm (Blei et al., 2003; Teh et al., 2006). And Xie (2019) proposed the social network analysis method, which can construct a scholar-paper-journal relationship network to identify potential author collaborations, journal publication trends, and research topics. Moreover, understanding the evolution of research topics is also an important research focus (He et al., 2009).

In this paper, the persona of CCSDY scholars is established based on the three attributes introduced above which are worthy of research and can depict scholars almost completely. Personal attributes include age and gender, educational attributes consist of the level of graduate institutions and academic ability of doctoral mentors, and academic attributes refer to the number of high-quality papers and research topics of high-quality papers obtained from Web of Science.

Method

Data collection

The data used in this paper were collected from the scholars who got funded by the CCSDY project from 1998 to 2019[iii], in all 80 scholars. For personal and educational attributes, the information of age, gender, graduate institutions and doctoral mentors of the CCSDY scholars was obtained from their online profiles on the university websites and various academic web databases. As for academic attributes, the title, publication date, journal or conference information of the papers that published by the scholars before they won the CCSDY project were extracted from the website AMiner, based on Scrapy technology. AMiner was used in this study as an academic database, which chronologically matches scholars to their papers from DBLP, ACM DL, CiteSeerX and other online data libraries (Wan, 2019). Once the data extraction was completed, the high-quality papers of CCSDY scholars were selected according to the academic conference and journal recommendation list released by the China Computer Federation (CCF)[iv]. Finally, to conduct the topic mining, the title of each high-quality paper was entered into the Web of Science search box. If the result existed, the abstract and keywords information in the web page would be parsed, and then Selenium technology was used to download them.

Curriculum vitae analysis method

Curriculum vitae analysis is a method for evaluating candidates by investigating their demographic information and work and education experience, which is often applied to research evaluation (Cañibano & Bozeman, 2009). Combining with statistical analysis methods and multi-aspect data visualization technology, curriculum vitae analysis is adopted in this paper to examine the personal and educational attributes.

Three-stage time series topic analysis model

The HDP algorithm, a hierarchical topic mining algorithm based on the Dirichlet process, is adopted as the core algorithm of the TTSTA model proposed in this paper. Unlike the LDA algorithm that requires users to preset the topic numbers, the HDP algorithm can automatically generate the number and content of the topics according to the characteristics of document sets (Teh et al., 2006). As an extension of the LDA algorithm with no parameters, the HDP algorithm is widely used in various text topic mining tasks (Zhou, 2011). The topic distribution of the document set in the HDP algorithm is represented by the probability distribution DP ($\boldsymbol{\alpha}$, H). For document J in a document set, its topic distribution G_j follows the distribution represented by DP (α_0, G_0) and the generation process of its topics is as follows: first determining the topic distribution of the document set by forming the Dirichlet process based on the base distribution H and parameter $\boldsymbol{\gamma}$, and then deducing the topic probability distribution $G_i \mid \alpha_0 G_i$ ~DP (α_0, G_0) of document J by forming the Dirichlet process based on the base distribution G_0 and hyperparameter $\boldsymbol{\alpha}$. The formula (1) for modelling the frequency of words in topics is shown below:

$$f(x_1, x_2 \ldots x_k, \alpha_1, \alpha_2 \ldots \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{a^i - 1} \tag{1}$$

B (α) in the formula (1) can be calculated by formula (2):

$$B(\alpha) = \frac{\prod_1^k \tau(\alpha_i)}{\tau(\sum_{i=1}^k \alpha_i)} \tag{2}$$

To analyse the research topics of CCSDY scholars in different periods and their topic evolution, the TTSTA model is proposed as shown in Figure 1. According to the publication date, all of the papers which can be found in Web of Science and published by each scholar before they got the CCSDY funding are sorted chronologically and in order to ensure the accuracy of the HDP algorithm and avoid the margin division error, these papers are equally divided into three stages: early, mid-term and recent. In case the number of papers can not be equally divided, the extra papers will be assigned to the stage where their closest published paper locates. Next, the HDP algorithm is used to obtain the research topics in each stage, and thus the common research topics and their evolution are finally acquired.

Personal Attribute Analysis

Age

The age of scholars can reflect the progress of relevant disciplines and the time patterns of scholar development. This study finds that the age of CCSDY scholars when they got funded was averagely 41 and ranged from 31 to 45. When they won the CCSDY project, only one of them was 31 years old, while there were 17 of them was 45 years old. Most CCSDY scholars were between 40 and 45 years old, and there were only five scholars younger than 35 years old. From 1998 to 2009, the average age of each year has generally shown an upward trend. And after 2009, the average age of each year was stably between 40 and 45.

These results can be explained by two possible reasons. Firstly, there is an age limit in the project requirement, where China stipulates that the applicants have to be under 45 years old on the first day of the application year. Therefore, scholars may strive harder when they were older than 40 years old to catch the last chance of being funded by the CCSDY project. Secondly, scholars need a long period to accumulate research achievements, cultivate academic sensitivity and judgment, and assemble excellent research teams. And it is unlikely that scholars who are new to the academic career can have such abilities and reach this stage.

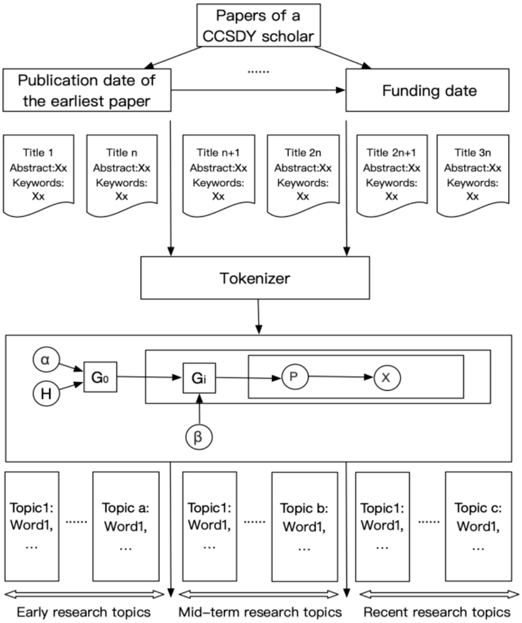

Papers of a CCSDY scholar
Publication date of the earliest paper
......
Funding date
Title 1 Abstract:Xx Keywords: Xx
Title n Abstract:Xx Keywords: Xx
Title n+1 Abstract:Xx Keywords: Xx
Title 2n Abstract:Xx Keywords: Xx
Title 2n+1 Abstract:Xx Keywords: Xx
Title 3n Abstract:Xx Keywords: Xx
Tokenizer
α
H
G0
Gi
P
X
β
Topic1: Word1, ...
Topic a: Word1, ...
Topic1: Word1, ...
Topic b: Word1, ...
Topic1: Word1, ...
Topic c: Word1, ...
Early research topics
Mid-term research topics
Recent research topics

Figure 1. Three-stage time series topic analysis model

Gender

Among 80 CCSDY scholars, there are five females (6.25%) and 75 males (93.75%), indicating a significant gender imbalance in the CCSDY project.

Such imbalance exists in various industries and research areas and can be attributed to many reasons. For instance, the traditional view that females are less good at engineering than males may cause a negative impact on female scholars in the computer science area (Smeding, 2019). Meanwhile, affected by the eastern traditional history and culture, female scholars in China often face more pressure than man when balancing family and scientific research, which also harms their long-term investment in the academic career (Xu, 2019).

Before 2018, only two female scholars got funded by the CCSDY project, and both of them had top doctoral mentors. One of them was supervised by a leading researcher of the Chinese storage field. And the other was supervised by an academician of Chinese Academy of Sciences, whose research on the *Goldbach conjecture* is widely acclaimed by the worldwide mathematicians (Wikipedia, 2020). Nevertheless, from 2018 to 2019, three female scholars accepted CCSDY funding, making positive progress on this issue.

Educational Attribute Analysis

Educational experience

After personal attribute analysis, the level of institutions the CCSDY scholars graduated from will be analysed in this section, including their bachelor, master and doctoral degrees. In China, universities can be ranked by academic level from highest to lowest in the following order: 985 universities, 211 universities, ordinary universities, and junior universities, and there are also some universities in Hong Kong and Macao. Other than universities, the Chinese Academy of Sciences (CAS) is also a top scientific research institute that can grant an academic degree in China. Three pyramids of the graduate institutions that cultivated more than two CCSDY scholars are presented in Figure 2.

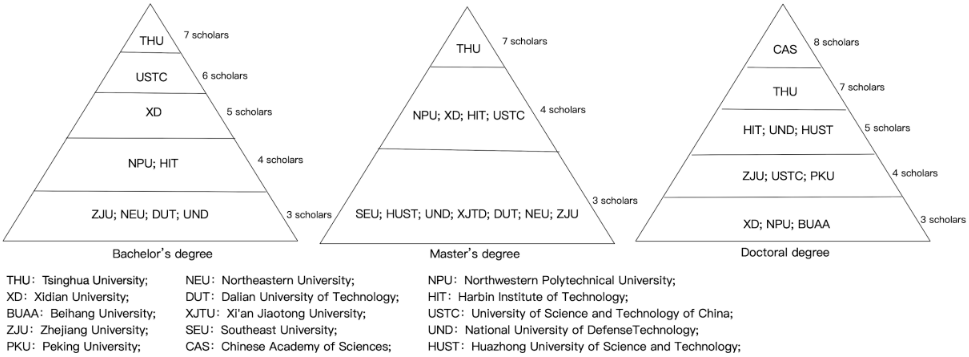

THU
7 scholars
USTC
6 scholars
XD
5 scholars
NPU; HIT
4 scholars
ZJU; NEU; DUT; UND
3 scholars
Bachelor's degree
THU
7 scholars
NPU; XD; HIT; USTC
4 scholars
SEU; HUST; UND; XJTD; DUT; NEU; ZJU
3 scholars
Master's degree
CAS
8 scholars
THU
7 scholars
HIT; UND; HUST
5 scholars
ZJU; USTC; PKU
4 scholars
XD; NPU; BUAA
3 scholars
Doctoral degree
THU：Tsinghua University;
XD：Xidian University;
BUAA：Beihang University;
ZJU：Zhejiang University;
PKU：Peking University;
NEU：Northeastern University;
DUT：Dalian University of Technology;
XJTU：Xi'an Jiaotong University;
SEU：Southeast University;
CAS：Chinese Academy of Sciences;
NPU：Northwestern Polytechnical University;
HIT：Harbin Institute of Technology;
USTC：University of Science and Technology of China;
UND：National University of DefenseTechnology;
HUST：Huazhong University of Science and Technology;

Figure 2. Pyramids of the graduate institutions of the CCSDY scholars

It is found that most CCSDY scholars completed their academic degree programs in 985 universities, among which the proportion of universities of science and engineering is significantly higher than comprehensive universities. This phenomenon may due to the fact that 985 universities possess relatively richer and better resources than other universities (Huang, 2014). Plus, universities of science and engineering focus more on practice than comprehensive universities, which better meets the cultivation need of CCSDY scholars owing to the practical features of computer science. Among the top part of the pyramids, Xidian University is the only 211 university, while the rest are 985 universities. Xidian University has long focused on electronics and information technology, which may indicate that an advanced level of certain disciplines can facilitate the development of scholars in relevant areas.

Tsinghua University is at the top of all the three pyramids. As a leading university in China, Tsinghua University has a strong computer science faculty and vast laboratory resources, contributing numerous outstanding scholars. Note that CAS joins the topmost part of the doctoral pyramid. In CAS, researchers identify themselves more as research staff instead of students, leading to different cultivation system from universities. And such enterprise-oriented education may have a positive incentive effect on the development of CCSDY scholars.

From left to right, Figure 3 reveals the path of the graduate institution level of CCSDY scholars, including bachelor's, master's, and doctoral degrees. Studying in 985 universities for the three academic degrees is the most common path among CCSDY scholars. Thirty-nine CCSDY scholars followed this path and got access to top mentorship and sufficient laboratory resources starting from their bachelor program. Besides, the education path of CCSDY scholars who completed bachelor programs in 211 or ordinary universities shows an upward curve in Figure 3, which means they tended to turn to research institutions at a higher level for better educational resources. However, it is found that CCSDY scholars who enrolled in Xidian University for their bachelor's degree prefer to continue studying in Xidian University instead of applying for a 985 university, which further manifests the advantage of Xidian University in cultivating computer science scholars.

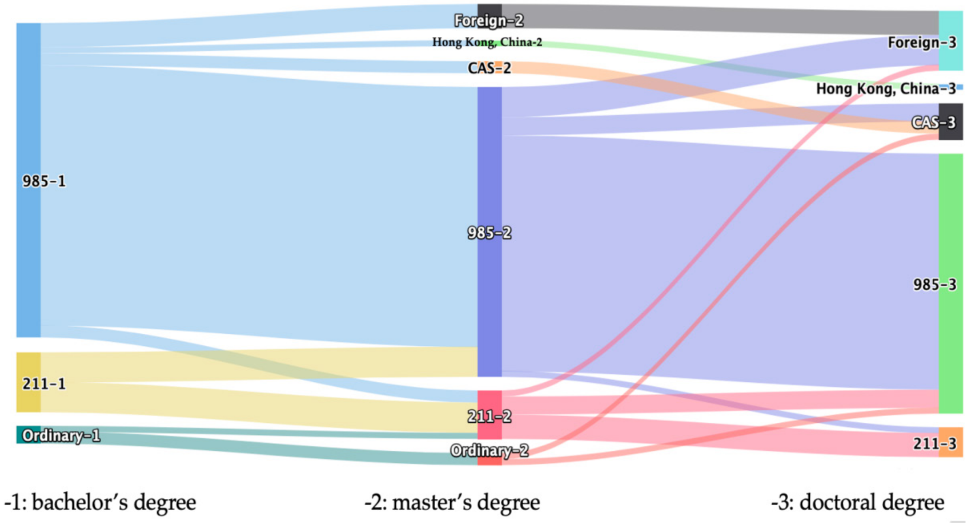


Figure 3. Higher education tracks of CCSDY Scholars

Furthermore, there has been a growing trend for CCSDY scholars to study abroad. Studying abroad allows scholars to absorb knowledge under superb guidance and broaden academic visions, which is an efficient way to improve creativity, develop new insights and help scholars make research breakthroughs.

Interaction and selection between research institutions and scholars. It can be seen from the educational experience that there exist two-way interaction and selection between universities and scholars. On the one hand, scholars prefer research institutions with more outstanding talents, advanced equipment and excellent mentors. On the other hand, academic institutions also tend to recruit distinguished scholars. Although it can not be denied that scholars already had high research potential before entering universities, the fact that most CCSDY scholars graduated from 985 universities shows a strong effect of research environment on the development of scholars.

As only five CCSDY scholars were younger than 35 years old when they received the funding, their curricula vitae were analysed in detail. Three of them worked as researchers in the top Internet companies in the United States, which helped them acquire practical experience and apply enterprise problem-solving skills to their research.

Doctoral mentorship

Mentorship is also a crucial factor in educational attributes. Based on the CCF field classification[v], this paper divides research fields of CCSDY scholars into seven: multimedia, network, architecture, software, data mining, cross and information security. The most popular research field is multimedia, which attracted 24 CCSDY scholars to contribute to it. After classifying CCSDY scholars into the seven research fields, their PhD mentors are divided into the corresponding fields. As shown in Figure 4, this paper classifies scholars with excellent academic ability into the following five categories: Academician of Chinese Academy of Science (ACAS), Academician of Chinese Academy of Engineering (ACAE), State Council Special Allowance Expert (SCSAE), the CCSDY scholars (CCSDY-mentor), and other top scholars. The scholars who do not belong to any of the five categories are called ordinary scholars in this paper.

Top mentors. It is found that two-thirds of the CCSDY scholars were guided by top mentors in their doctoral program. And the proportions of top mentors in software (77.78%), architecture (90.91%), and cross (77.78%) are relatively large. Top mentors can help students develop research interests, proper research habits and critical thinking skills. At the same time, students can use their mentors' personal networks to gain better resource. For example, they can know more outstanding scholars with whom they can communicate and get excellent inspiration. However, many scholars can still become outstanding without top mentors, especially in applied fields such as network (33.33%) and information security (40.00%). Because these applied fields involve a wide range of research objects and applications and are in a vigorous development period, scholars can independently exert their research strength on a specific issue in these fields. Interestingly, a leading scholar in architecture cultivated two CCSDY scholars, and one of them also supervised one CCSDY scholar, presenting a clear line of the mentoring relationship.

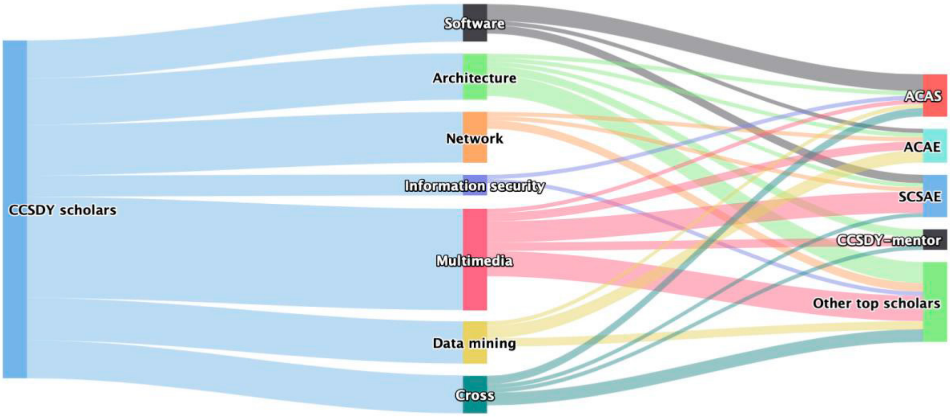


Figure 4. Mentor's field and academic ability classification

Academic attributes analysis

High-quality papers

A bibliometric analysis was conducted on high-quality papers published by scholars before being funded by CCSDY project. The average number of high-quality papers published per year shows a gradual growing trend from 1998 to 2019, which rose rapidly in 2015. On average, each CCSDY scholar published 63 high-quality papers before they received the funding. In recent years, CCSDY scholars have published an increasing number of papers on international journals and conferences, sharing research achievements with worldwide researchers. Following the lead of CCSDY scholars, there will be a growing trend of Chinese computer science scholars to be active on the international stage and contribute to the whole world.

Research topics

Although the research of CCSDY scholars may not be the most representable in computer science area, it is useful and meaningful for the country's development, which can help other scholars for reference. The research topics of CCSDY scholars are examined by the proposed TTSTA model from both vertical and horizontal perspectives. Vertical perspective refers to discovering common research topics among multiple scholars and focuses on the research hotspots. Horizontal perspective means to analyse the evolution of the research topics and focuses on the academic development of one single scholar. Due to the enormous quantity of data, only the research topics of 3326 papers from 37 scholars who got funded in the last five years are analysed, and only the research topics of four fields are displayed in Table 1 and discussed in the analysis below.

Table 1. The samples of CCSDY scholars' research topics in different fields.

Fields in computer science	*Early Topics*	*Mid-term topics*	*Recent topics*
Multimedia	Face detection/recognition Facial expression recognition Image segmentation ...	Facial expression analysis Expression recognition Image retrieval ...	Activity identification in group Object segmentation Facial expression analysis ...
Data mining	Keyword extraction Web topic discovery Citation recommendation ...	Biological relationship mining Academic data analysis Name disambiguation ...	Recommended friends on Social networks Topic extraction Cross-language knowledge Base construction ...
Network	Network data packet transmission method Data packet Network data packet transmission efficiency ...	P2P network Packet compression Route lookup ...	Video delivery Network attack detection Network transmission loss data recovery ...
Software	Software component configuration Component management Web container ...	Web application overload control Version compatibility of software evolution Location and recovery of failed components ...	Open-source project development and maintenance Classification of problems in software development Analysis of community members in open source software projects ...

Firstly, in the multimedia field, scholars all contribute to applying technology to practical scenarios. Most multimedia scholars focus on the classification, positioning, detection and segmentation of information in videos and images. In its sub-field of video research, the analysis of the movement in videos and the video plagiarism detection are the research focus. In another sub-field, image research, the most popular topics are face recognition and facial expression recognition, which have high commercial and practical value. As for the topic evolution, from face recognition, facial expression recognition to micro-expression recognition, it can be seen that the research in this field is more refined and the technical complexity becomes higher, which can bring more accurate results. For example, micro-expression recognition can be used to determine the change of people's emotion, and thus can be utilised in the medical field for mental treatment or in the national security field for deception detection (Takalkar et al., 2018). All in all, such refinement in research can help lead to a broad prospect of practical application.

Secondly, CCSDY scholars in the field of software are most interested in web services and software development. The former includes evaluation, optimization, search and recommendation of web services. The latter consists of component configuration, combination, management, faulty component location and recovery. Besides, it is discovered that the evolution of research topics in software development is highly consistent with real-life technical needs. Thirdly, in the data mining field, one of the research paradigms is first conducting topic mining and keyword extraction, based on which the later research will explore new topics with prospects of technical integration and practical application. Another paradigm is to first focus on practical issues in specific fields, and continuously deepen and refine the research topics. Finally, the network field mainly focuses on basic research subjects, for example, routing lookups of data transmission, and the avoidance and detection of network congestion.

Three topic evolution paths. From the changes of research topics in different periods, it can be concluded that there are three paths of research topic evolution among the CCSDY scholars: (a) from basic technical problem to practical applied problem; (b) from existing technology to new integrated technology; and (c) from macroscopic topics to in-depth topics.

The persona of CCSDY scholars

After analysing the common characteristics of CCSDY scholars through three attributes, the persona of CCSDY scholars is established as shown in Figure 5.

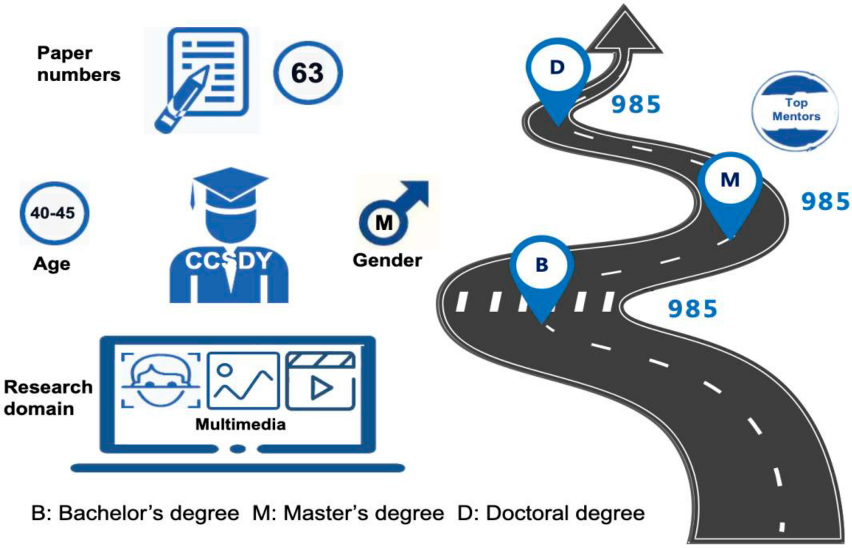


Figure 5. The persona of the CCSDY scholars

From this persona, the image of CCSDY scholars can be interpreted as follows:

1. Most CCSDY scholars are between 40 and 45 years old, and there are far more male scholars than women scholars.
2. As for higher education, CCSDY scholars tend to study in 985 universities from bachelor to doctorate and be supervised by top mentors in their doctoral study.
3. The average number of high-quality papers published by the CCSDY scholars is 63, showing an upward trend. And many of the CCSDY scholars mainly work in image and video research, in which face recognition, image segmentation and video detection are the most popular research topics.

Discussion and conclusion

This paper has analysed common characteristics of the CCSDY scholars and completed a persona of them, based on which the following conclusions and practical implication on scholar cultivation in the computer science area are drawn.

To begin with, the ageing and gender imbalance of CCSDY scholars should raise the attention of the funding institutions. The energy and passion of young scholars are invaluable, and thus the relevant preferential policies to promote young scholars are expected to be introduced. This paper has also shown the lack of female CCSDY scholars, suggesting that more support for females in education and research funding are needed to ensure a fair environment for them to study and work.

By analysing the work experience of relatively young CCSDY scholars, this paper has found that most of them worked as researchers in American top Internet companies. The research experience in an enterprise can enhance coding ability and understanding of practical applications, which improves both problem-finding and problem-solving ability. Therefore, more school-enterprise cooperative research projects will be more conducive to the development of computer science scholars.

Through the analysis of educational experience, it is found that there are interaction and selection between research institutions and scholars. It is more beneficial for scholars to apply for a higher-level institution, and capable universities are inclined to recruit excellent students. Also, the top research strength and reputation of high-level institutions are of great help for scholars to win the funding project. And studying in universities of science and engineering are advantageous for the development of scholars in the computer science area.

Next, the guidance from top mentors is a strong positive incentive for CCSDY scholars. Two-thirds of CCSDY scholars were supervised by top mentors, which caused the Matthew effect of distinguished scholars. If the country makes more efforts to cultivate the CCSDY scholars, they will also bring more outstanding students and core scientific research outputs. The dependence of top mentors varies in different fields. Scholars in applied fields such as network and information security are not reliant upon top mentors. For the fields of architecture, software, and cross, the proportion of top mentors is relatively large.

Besides, the distribution of CCSDY scholars in different computer science fields is also unbalanced, and there are far more CCSDY scholars in multimedia than in other fields. It is crucial to develop hot research fields, but other fields also need advancement. Moreover, basic research is the foundation of technological development and insufficient support for basic research may cause stagnation after a short period of boom. Therefore, academic funding projects should pay more attention to less popular fields and basic research fields.

Finally, there are three main characteristics of the research topics of CCSDY scholars: (a) in-depth exploration, which means refining research questions; (b) interdisciplinary expansion, which means extending and applying existing technology to new research objects and fields; and (c) integration, which means the combination of existing technology and new technology. Based on consolidated basic research, promoting communication among scholars within the field can facilitate integrating new technologies. Sharing of cross-disciplinary hot topics, technical exchanges, and research collaboration will also benefit the interdisciplinary expansion.

Limitation

This paper has conducted an in-depth analysis of factors in talent development and provides a deeper insight into the cultivation of computer science scholars in China. Nevertheless, there are still several limitations and recommendations explained as follows:

- Family and work experience are essential elements in the growth of scholars. But this paper is limited by the lack of these information attributed to the difficulty of relevant data collection. A further study could assess the effects of family and work experience factors.

- There are correlation and interactions between multiple subjective or objective factors in the development of scholars. Unfortunately, the paper did not examine the relationships between these factors, which would be a fruitful area for further work.
- Since there is no public information about the scholars who applied to *the National Science Fund for Distinguished Young Scholars* project but failed, it is difficult to explore the correlation between funding and the three characteristic groups.
- In the research topics section, because the sources of all high-quality papers are varied, abstracts and keywords are difficult to get, so we only use the papers which come from Web of Science. If we can analyse all high-quality papers, this research will be more complete.
- The data used in this paper were obtained from several open web databases, and there is a small part of missing data, which may lead to slight deviations in the analysis.

Acknowledgments

We would like to thank Qiqi Wang for her extraordinary support in the writing process. And this work was supported in part by the Natural Science Foundation of China under Grant 61976036.

References

Archer, S. L. (2007). The making of a physician-scientist—the process has a pattern: Lessons from the lives of Nobel laureates in medicine and physiology. *European heart journal*, 28(4), 510-514.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3, 993-1022.

Cañibano, C., & Bozeman, B. (2009). Curriculum vitae method in science policy and research evaluation: The state-of-the-art. *Research Evaluation*, 18(2), 86-94.

Chambers, E. G., Foulon, M., Handfield-Jones, H., Hankin, S. M., & Michaels III, E. G. (1998). The war for talent. *The McKinsey Quarterly,* (3), 44.

Frosch, K. H. (2011). Workforce age and innovation: A literature survey. *International journal of management reviews*, 13(4), 414-430.

Guo, A., & Ma, J. (2018). Archetype-based modeling of persona for comprehensive personality computing from personal big data. *Sensors*, 18(3), 684.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009). Detecting topic evolution in scientific literature: How can citations help?. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 957-966).

Huang, Y. H.(2014). Analysis on Differences of Library Resources Allocation between “985” University and Ordinary University. Information Research(09),66-69+73.

Junior, P. T. A., & Filgueiras, L. V. L. (2005). User modeling with personas. In *Proceedings of the 2005 Latin American conference on Human-computer interaction* (pp. 277-282).

Johansson, M., & Messeter, J. (2005). Present-ing the user: Constructing the persona. *Digital Creativity*, 16(04), 231-243.

Kong, X., Jiang, H., Yang, Z., Xu, Z., Xia, F., & Tolba, A. (2016). Exploiting publication contents and collaboration networks for collaborator recommendation. *PloS one*, 11(2), e0148492.

Lin, W. (2017, August). Analysis of the lack of scientific and technological talents of high-level women in China. In *IOP Conference Series: Earth and Environmental Science* (Vol. 81, No. 1, p. 012044). IOP Publishing.

Lu, W. J., Xu, L., Liu, J. & Chen, J. (2018). Review on Chinese Artificial Intelligence Research in Past Decade: Bibliometric of Knowledge Atlas during 2008-2017. *Journal of Technology Economics* (10), 73-78+116.

Margot, K. C., & Kettler, T. (2019). Teachers’ perception of STEM integration and education: A systematic literature review. *International Journal of STEM Education*, 6(1), 1-16.

National natural science foundation of China. (2018). 2019 Manual of National Science Fund for Distinguished Young Scholars project. Retrieved from http://www.nsfc.gov.cn/nsfc/cen/xmzn/2019xmzn/08/index.html

Pruitt, J., & Grudin, J. (2003, June). Personas: Practice and theory. In *Proceedings of the 2003 conference on Designing for user experiences* (pp. 1-15).

Shen, C. C., Hu, Y. H., Lin, W. C., Tsai, C. F., & Ke, S. W. (2016). Research impact of general and funded papers. *Online Information Review.* 40(4), 472-480.

Smeding, A. (2012). Women in science, technology, engineering, and mathematics (STEM): An investigation of their implicit gender stereotypes and stereotypes' connectedness to math performance. *Sex roles*, 67(11), 617-629.

Takalkar, M., Xu, M., Wu, Q., & Chaczko, Z. (2018). A survey: facial micro-expression recognition. *Multimedia Tools and Applications*, 77(15), 19301-19325.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American statistical association*, 101(476), 1566-1581.

Wan H. Y., Zhang Y. T., Zhang J., & Tang J. (2019). AMiner: Search and Mining of Academic Social Networks. *Data Intelligence*, 1(1), 58-76.

Wikipedia. (2020). Pan Chengdong. In *Wikipedia*, *The Free Encyclopedia*. Retrieved from https://en.wikipedia.org/wiki/Pan_Chengdong

Xie, P. (2019).Recognition of Scholar Interest Tag for Academic Literatures. *Technology Intelligence Engineering*, 5(03), 65-73.

Xu, J. (2019). The plights of personal development encountered by female scientists in China. *Journal of Xiangyang Polytechnic*. 18(06), 78-83.

Yu, X., Wei, H. N., Wang, X. L., & Wang X. F. (2020). Analysis of National Science Fund for Distinguished Young Scholars for promoting discipline layout development of basic research. *Bulletin of National Natural Science Foundation of China*, 34(05), 621-629.

Zhao, S. X., Lou, W., Tan, A. M., & Yu, S. (2018). Do funded papers attract more usage?. *Scientometrics*, 115(1), 153-168.

Zhou, J. Y., Wang, F. Y., & Zeng D. J. (2011). Hierarchical Dirichlet processes and their applications: A survey. *Acta Automatica Sinica*, 37(04), 389-407.

Zhu, J. (2019). The composition and evolution of China's high-level talent programs in higher education. *ECNU Review of Education*, 2(1), 104-110.

[i] Yuan Lin, Xinhang Zhao and Jiaojiao Feng contribute equally to this work.

[ii] Artificial intelligence is a branch of computer science, so for the convenience of later explanation, in this paper, "the computer science area" refers to the combination of computer science and artificial intelligence. The research areas chosen in this study are also defined by NSFC.

[iii] The *National Science Fund for Distinguished Young Scholars* project began in 1994, but the area of computer science started in 1998. And the data in 2019 are the newest available data.

[iv] CCF recommended international academic conference and journal directory can be seen at https://www.ccf.org.cn/c/2019-04-25/663625.shtml.

[v] The seven fields of computer science area are defined in this paper based on the classification of conference and journal by CCF.

A retrospective analysis of China’s science and technology evaluation policy since 1978

Xin Liu [1], Zhenyue Zhao[2] and Jiang Li [3]

[1] *liuxin@smail.nju.edu.cn*
School of Information Management, Nanjing University, Nanjing (China)

[2] *njujack@163.com*
School of Information Management, Nanjing University, Nanjing (China)

[3] *lijiang@nju.edu.cn*
School of Information Management, Nanjing University, Nanjing (China)

Abstract

China’s science and technology evaluation policy (S&TEP) guides China’s science development. The purpose of this research is to analyze the topic evolution of China’s S&TEP and the public’s response to the S&TEP. We take reform and opening up in 1978 as the starting point and divide the development process into four stages. We collected 170 S&TEP retrieved from database. Latent Dirichlet Allocation (LDA) modeling was used to identify policy topics of each stage. Then, taking the policy *Suggestions on the Correct Use of SCI as Research Evaluation Criterion and the Establishment of Proper Guidelines for Research Evaluation (2020)* as an example, the public understanding of this policy was summarized by analyzing comment data from Chinese newspapers, academic papers, social media, and foreign media websites. The results show that: (1) most significantly, the topic of China’s S&TEP has shifted from tilting to the quantity of publications to raising the quality of publications; (2) the evaluation approach has shifted from peer review to quantitative evaluation, and to breaking quantitative evaluation and building a comprehensive system at present; and (3) the text of China’s S&TEP has the implications beyond the text itself, but they are easy to reach common acceptance among Chinese scholars.

Introduction

Since the implementation of the reform and opening-up in 1978, China has made significant progress in science and technology innovation. Science and technology (hereafter referred to as S&T) evaluation is an important part of S&T management. Looking back at the changes in China’s science and technology evaluation policy (hereafter referred to as S&TEP) over the past four decades, they are often accompanied by major problems encountered in S&T management. In early 2020, the Ministry of Education and the Ministry of Science and Technology jointly issued two policies on discouraging the use of SCI and paper-oriented evaluation system, which marks the beginning of the reform of China’s S&T evaluation system. Such a revolutionary transition keeps drawing serious attention of the entire S&T community in China and even abroad from the very moment they were brought about (Mallapaty, 2020). Discussions on them appear on newspapers, journals, online forums, social media, etc. However, the discussion appears to be highly controversial and non-systematic. Some questioned the rationality and viability of specific items in the policy documents or the policies as a whole (Li, 2020), while others concerned what are to be the adoptions over the renunciations (Chen & Zhang, 2020; Ye, 2020). Still others seek to reveal the “real intentions” of the policy makers, evaluating the situation that China’s science and technology has been in face of (Qian et al.,2020).

Our study aims to analyze the topic trends of China’s S&TEP and public response to the new policy, and reveals the evolution pattern and inner meaning of China's S&TEP to help researchers better understand policies and provide insights for the future adjustment of S&T evaluation systems. The study is divided into two parts. In the first part, we investigate the development history of China’s S&TEP and grasps the characteristics of each stage by analyzing the topic evolution of policies at different stages. The second part is presented as a case study, in which we take the policy about reforming SCI-oriented evaluation system as an example and conduct a systematic investigation of the public discussions on the policy.

Data and methodology

Data collection

The data set includes S&T policy documents and the data of public comments. The policy documents are obtained from PKULAW (A Database of China Policies developed by Peking University). In order to cover all the important policies related to China's S&T evaluation formulated at the national level since 1978, we conducted full-text retrieval with search terms such as "research evaluation" and "S&T evaluation". Due to the large number and diverse content of the policies, we manually sorted the policies according to the following criteria to ensure the accuracy of sample selection: (1) the effectiveness of the policy is administrative regulations and departmental rules, which were issued by the CPC Central Committee, the State Council and related departments; (2) the type of policies does not include leadership speeches, letters, instructions, etc.; and (3) the content of the policies is closely related to S&T evaluation. After repeated selection and verification, 170 effective policies were finally sorted out. Figure 1 shows changes in the number of policies over time.

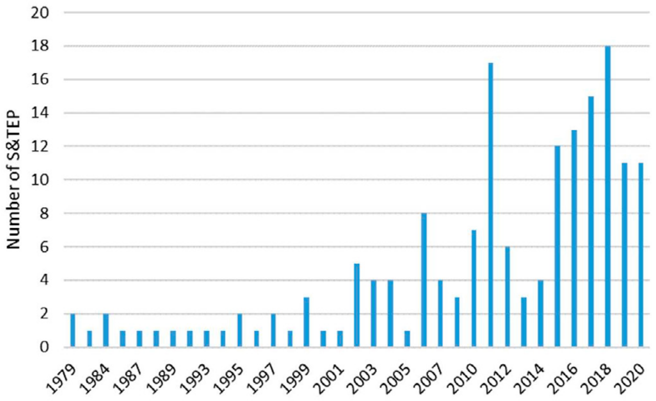


Figure 1. The annual number of China's S&TEP since 1978.

The public comments data set is about the policy *Suggestions on the Correct Use of SCI as Research Evaluation Criterion and the Establishment of Proper Guidelines for Research Evaluation*. Since the S&TEP is mainly for S&T workers, the data we selected come from four channels: domestic official media, academic online forum, scientific papers, and foreign media. Among them, domestic official media data are obtained from two important Chinese newspapers--the People's Daily and Guangming Daily. They are authoritative sources for obtaining official news, policy propositions, and mainstream ideas in China. The academic online forum comment data is collected from the ScienceNet website (http://www.sciencenet.cn/). ScienceNet is currently the world's largest Chinese science community and an important platform for Chinese scholars to conduct academic exchanges. The relevant data of scientific papers is obtained from CNKI, which is a Chinese academic journal database. The number of policy comment data for various channels is shown in the Table 1.

Table 1. The number of comment text in each source.

channel	*source*	*number*
Newspaper	People’s Daily & Guangming Daily	35
Online forum	ScienceNet	48
Scientific papers	CNKI	59
Foreign media	GOOGLE	15

Methodology

For sorting out the development process of S&TEP, we adopted the LDA topic model method. LDA (Latent Dirichlet Allocation), is an important text mining method used in topic discovery tasks to reveal hidden topic patterns in large-scale corpora. The model is developed from pLSA (Hofmann, 1999), which solves problems of pLSA such as overfitting (Blei et al., 2003). Since it is an unsupervised learning method that does not require labeled training set samples, it is one of the most widely used models today.

LDA is essentially a three-level hierarchical Bayesian model. It uses “bag-of-words” model to represent documents and maps complex high-dimensional words in unstructured documents into a low-dimensional space of "document-topic-word" to obtain document clustering. The basic idea is that the documents in corpora are represented as the probability distribution of a series of latent topics, and each topic is represented as a series of word distributions (Blei et al., 2003).

In this study, LDA was performed by using Gensim Python library. Before modeling, some preprocessing work has been done on the policy texts such as eliminating stop words. Then we apply the LDA model to the full text of the policy at each of the four stages. Ten keywords with the highest probability are selected to represent the characteristics of each topic. For each policy text, the topics are ranked according to probability values, and the top 3 categories are selected as the most representative topics. Through the comparison of the probability distribution of the topic, the evolutionary characteristics of China’s S&TEP are discussed.

How to determinate the best number of topics is an important but controversial matter (Arun et al. 2010). Perplexity is a topic clustering performance evaluation method based on a small data set, which is often used as a criterion to determine the number of topics (Blei et al., 2003; Miyata et al., 2020). It refers to the degree of uncertainty of the model about which topic the document d belongs to. We draw the perplexity curve after trying different topic numbers, and use the pyLDAvis tool to visualize the generated topics. The appropriate number of topics can be found by combining the visualization results and perplexity values.

Analysis of the development stage of China’s S&T evaluation system

Scholars have different opinions on how to classify the development stages of China’s S&T evaluation system. Xu et al. (2018) divided the development history into three stages: restoration and reconstruction(1978-1984), exploration and practice(1985-2003), and institutional regulation and construction(2003-); Gou et al. (2020) analyze the development of research evaluation policy in China and divided it into four stages: initial development period (1949-1976), restoration and adjustment period(1977-1989), wandering change period(1990-2001), and perfection and advancement period(2002-); through the analysis of science and technology evaluation papers, Du et al. (2020) divided China’s S&T evaluation research into four stages: 1978-1993, 1994-1997, 1998-2007, 2011-2015. In general, scholars’ division of the stages of S&T evaluation system development is mostly based on important national historical events or strategic shifts, but the analysis of policy development after 2000 is relatively weak. In order to fill the gap, based on the reform of S&T system and the release of important policies in S&T evaluation activities, we explore new division of the development of

S&T evaluation in China since 1978. The evolutionary history can be divided into four periods, which have their own characteristics.

1978-2003

We conducted a topic analysis of the 29 policy documents in this phase, and the keywords and their corresponding probabilities of each topic are shown in Table 2. Talent development (A1), S&T awards (A2), and research management (A3) were the key topics in this stage.

Table 2. Latent Dirichlet Allocation results for the period 1978–2003.

Topic	*Keyword*	*PR*	*Keyword*	*PR*
Talent Development (A1)	S&T Evaluation	0.007	Life Benefits	0.0023
	S&T	0.0044	S&T	0.0023
	Human Resources	0.004	Jury	0.0022
	Declaration	0.0037	Technical Staff	0.0019
	Medical	0.0025	Young and Middle-aged	0.0019
S&T Awards (A2)	Progress Award	0.0047	Change	0.0032
	Candidates	0.0044	Teachers	0.0029
	S&T Award	0.0044	Grant	0.0027
	Academic Ethics	0.0034	Award	0.0027
	Evaluation	0.0032	Committee	0.0021
Research Management (A3)	Colleges and Universities	0.0049	Research Institutions	0.003
	Assessment	0.0032	Review Experts	0.0027
	Humanities & Social Science	0.0032	Change	0.0024
	Program	0.0031	Training Program	0.0022
	Basic Research	0.0031	Originality	0.0022

The most recent revolutionary shift in China's strategy of S&T development can be dated back to the late 1970s. In March 1978, the first National Congress of Science was held in Beijing (Wang, 2018). The congress sought to rectify the attitude of government and the people towards S&T as well as S&T workers. These events signalled that China has entered a new phase in the development of Science and Technology. Since then, China has successively restored its professional title system and S&T award system. Topics A1 and A3 of policy in this stage also illustrate the point. In 1985, China began to reform the S&T evaluation system, proposed to break the long-term planned economic system, and gradually increased funding for various competing projects (Wang et al., 2018). This reform marked the beginning of project funding from a government-led model to a competitive model. Project establishment and review, S&T achievements evaluation, and performance evaluation have become important parts in scientific and technological activities.

The reforms have largely reshaped the S&T community in China. The urgency of compatible and efficient evaluation systems was uttering. Before the 1980s, the evaluation system in China was generally criticized as highly subjective which is mainly based on the opinions of experts and administrators of institutes. The peer-review-based evaluations without well-designed instructions and restrictions were not proved to be well-functioning in the society where *guanxi* (a Chinese word representing the complex social relationships between individuals) played a significant role (Zweig, 1997; Zweig, 2006; Zweig & Wang, 2013). The introduction of foreign bibliographic databases such as SCI into China's evaluation systems in the 1980s started an era of "quantity". In 1987, the Institute of Scientific and Technical Information of China (ISTIC) first used the numbers of publications indexed in several bibliographic databases to evaluate the research ability of institutes in China. This is marked as the first formal attempt in China to use quantitative methods based on bibliographic indexes in the evaluations. For the evaluation of individuals, Nanjing University was the first to build direct links between the number of SCI papers and monetary awards (Shao & Shen, 2012). Owing to such a policy, Nanjing University produced the largest number of SCI papers among all the universities in China for 7 straight years in the 1990s. Some argue that they have opened the "Pandora's box" because the usage of bibliographic indexes spread quickly to the whole country and eventually turned the research evaluation systems into SCI-based and SSCI-based ones (Cao, 2008; Quan et al., 2017)

2003-2010

We conducted a topic analysis of the 24 policy documents at this stage, and the results are shown in Table 3. The topics of the second-stage can be sorted into three categories: research integrity (B1), research evaluation (B2), and talent evaluation in various fields (B3).

Table 3. Latent Dirichlet allocation results for the period 2003–2010.

Topic	*Keyword*	*PR*	*Keyword*	*PR*
Research Integrity (B1)	Academic Misconduct	0.0074	Institution	0.0037
	Investigation	0.0072	Research Integrity	0.0032
	Scientific Research	0.005	System	0.003
	Project	0.0049	Informant	0.0026
	Professional	0.0044	Review	0.0026
Research Evaluation (B2)	Evaluation	0.0098	Innovation	0.0046
	S&T	0.006	Research	0.0043
	Philosophy & Social Sciences	0.0056	Work	0.0042
	National	0.0049	Management	0.0032
	S&T Plan	0.0049	Colleges & Universities	0.0029
Talent Evaluation in Various Fields (B3)	Talents	0.0098	Research	0.0046
	Technologists	0.0072	Science	0.0043
	S&T	0.0062	Medical & Health	0.0041
	Rural	0.0048	Academic	0.004
	Academic Ethics	0.0046	Evaluation	0.0035

In the 1990s, the introduction of SCI and the promotion of quantitative indicators such as the number of papers caused S&T evaluation activities to heavily rely on quantitative indicators (Xu et al.,2018). The practice of "judging heroes by quantity" and the linking of S&T evaluation results with profits have led to scientific researchers' eagerness for quick success and profit. At the same time, the problem of "guanxi" of peer review in the evaluation activities such as project, job title and achievement have still not been well resolved. In addition, the development of the scientific research evaluation system is not perfect. The evaluation of research results in different disciplines and fields often adopts a unified evaluation standard. The irrationality of this approach has gradually been exposed. These problems began to arouse the attention of academia and the government. A very important topic of evaluation policy issued by government in this stage is research integrity (B1).

The year 2003 was an important turning point in the development of China's S&T evaluation. Five departments jointly issued the Decision on Improving the Work of Science and Technology Evaluation. The decision clarified that S&T evaluation should always put quality in the first place, pay attention to original innovation, and must not replace quality with quantity. For talents, different evaluation indicators should be determined according to different levels and types of scientific and technological work. Subsequently, in 2007, the NPC Standing Committee revised the Law of the People's Republic of China on Science and Technology Progress and formally established the legal status of the S&T evaluation system, pointing out that the state should "establish and improve a science and technology evaluation system that is conducive to funding innovation", and implement classified evaluation in accordance with the principles of fairness, justice and openness.

On the whole, China's S&T evaluation activities have gradually become formal and standardized at the stage. Various policies have emphasized two initiatives that focused on the quality of results rather than numbers and the implementation of categorical evaluation. The evaluation methods for talents are gradually standardized. As it can be seen in Topic B3, policies related to talent evaluation are beginning to be developed in areas such as agriculture and medicine and health.

2010-2018

We analysed 76 policy texts in this phase, and the results are shown in Table 4. Research evaluation in universities (C1), research evaluation in medical field (C2), S&T talent evaluation (C3) and talent evaluation in various field (C4) were the key topics in this stage.

Table 4. Latent Dirichlet Allocation results for the period 2010–2018.

Topic	*Keyword*	*PR*	*Keyword*	*PR*
Research Evaluation in Universities (C1)	Think Tanks	0.0054	Papers	0.0026
	Teachers	0.0037	Research & Development	0.0018
	S&T Workers	0.0032	S&T Innovation	0.0017
	Colleges & Universities	0.0032	Philosophy & Social Sciences	0.0016
	Titles	0.0027	Chinese Characteristics	0.0016
Research Evaluation in Medical Field (C2)	Public Hospitals	0.0035	Prosperity	0.002
	Medical Staff	0.0034	Clinical	0.002
	Expert Database	0.0029	Chinese Medicine	0.0019
	Academic Journals	0.0027	Medical and Health	0.0018
	S&T Innovation	0.0022	Research & Development	0.0018
S&T Talent Evaluation (C3)	S&T	0.0033	Fundamental	0.0023
	Talents	0.0032	High-Level	0.0022
	Postdoctoral	0.0031	Training	0.0019
	Women	0.003	Academic Atmosphere	0.0018
	Evaluation	0.0028	Incentive Mechanism	0.0017
Talent Evaluation in Various Fields (C4)	Chinese Medicine	0.0036	Intellectual Property	0.0025
	S&T Talents	0.0029	Talent Selection	0.0022
	Geomatics	0.0029	Training	0.0021
	Academic Journals	0.0026	Geographic Information	0.002
	Food	0.0025	Agriculture	0.0016

In 2010, the State Council successively issued the *Outline of the National Medium and Long-term Talent Development Plan (2010-2020)* and the *Outline of the National Medium and Long-term Education Reform and Development Plan (2010-2020)*. Subsequently, China's 12th Five-Year Plan was launched. *The outlines* propose to "improve the innovation and quality-oriented scientific research evaluation mechanism" and "overcome the tendency of only academic, only qualifications and only papers". The dissemination of the outline lays the foundation for the formulation of Research Evaluation in Universities and talent evaluation plans. Various fields have successively issued policies related to the evaluation of scientific and technological talents in this field. It can be reflected in the policy texts as Topic C3 and Topic C4.

In 2013, the Ministry of Education issued the *Suggestions on Deepening the Reform of Scientific and Technological Evaluation in Colleges and Universities*, pointing out the problems in S&T evaluations such as focus on quantity at the expense of quality and focusing on form and neglecting content. And *the Suggestion* emphasized the need for classified evaluation in colleges and universities, increasing international peer evaluation and third-party evaluation, focusing on the quality of technological innovation and practical contributions. In 2016, president Xi proposed at the National Science and Technology Innovation Conference to "establish a classification and evaluation system oriented on the quality, contribution, and performance of scientific and technological innovation", which further pointed out the direction for optimizing S&T evaluation.

2018-present

We conducted a topic analysis of 40 policy documents at this stage, and the results are shown in Table 5. The policy text at this stage can be summarized into research evaluation (D1), *Four only* (D2), and S&T innovation (D3).

Table 5. Latent Dirichlet Allocation results for the period 2003–2010.

Topic	*Keyword*	*PR*	*Keyword*	*PR*
Research Evaluation (D1)	Research Integrity	0.0048	Basic Research	0.0021
	Social Science	0.0028	Program	0.0021
	Education	0.0027	Jury Committee	0.0019
	Talent Title	0.0025	Research Institutes	0.0018
	Title Review	0.0023	Performance Evaluation	0.0017
"the Onlys" (D2)	Four Only	0.0057	Title Review	0.0033
	Clean Up	0.0043	Only Title	0.0027
	SCI Papers	0.004	Only Awards	0.0024
	S&T	0.0036	Violations	0.0023
	Related Indicators	0.0036	Only Qualifications	0.0019
S&T Innovation (D3)	Fundamental Research	0.0034	Students	0.0016
	Science Fund	0.0026	Academic Journals	0.0016
	Support	0.0022	Mechanism	0.0015
	R&D	0.0017	S&T Innovation	0.0015
	Papers	0.0016	Philosophy & Social Sciences	0.0014

In July 2018, the Chinese government has issued a policy aimed at reforming the evaluation systems of projects, workers and institutes in the S&T section. For the evaluation of workers in the S&T section, specifically, the policy document prescribes that "such tendency must be renounced that the evaluation for individuals based solely on their academic papers, solely on their professional titles, solely on their educational backgrounds or solely on their received awards and honours". A formulation later widely referred to as *four only* thus came into being. Later in November 2018, the *four only* formulation was extended into *five only* by the Chinese Ministry of Education who then launched a special action aimed at implementing such policies in universities. Behind the simplistic formulations are a series of principles, rules, restrictions and ideas that seek to govern China's S&T evaluation practices in all aspects. For simplicity, in the below text, we will refer to the policies related to the *four only* or the *five only* altogether as *the only*. Topic D2 reflects this focus of work during the stage.

The ultimate goal of *the only* is, as is prescribed by the Ministry of Education, to reform China's S&T evaluation systems into quality-oriented ones which mainly consider the quality, contribution and impact of works. It is also proposed that the evaluation be based on the representative works, i.e., a subset of all the works, of the subject, thus breaking the quantity-oriented traditions to a larger extent. This main purposes of *the only* can be broken down into two. The first is that the evaluation activities not be over-simplified so that a single evaluation item, which could be academic paper, professional titles, etc., becomes of paramount importance and decisive for the evaluation. Another intention is to standardize the use of foreign bibliographic indexes such as SCI, SSCI and EI. These bibliographic indexes have greatly re-shaped the evaluation systems in a way that they became the only criteria to judge the quality of research outputs (Cao, 2008; Quan et al., 2017). Such an orientation has indeed helped China's S&T earn higher impact and a prestigious position in the global society over the past few decades. However, as they no longer fit China's current quality-oriented evaluation systems, they are to be criticized and renounced.

The *only* is, of course, not merely advocations. The Chinese governments have shown great resolution through a series of policies and actions that followed up. In February 2020, for example, two back-to-back policies were issued by the government aiming specifically at renouncing the mis- and over-using of academic papers in the evaluations. The former, issued by the Ministry of Science and Technology, prescribed several punishments for institutes that refuse to practice their evaluation activities in accordance with the policies. The latter, issued by the Ministry of Education and the Ministry of Science and Technology, was more particular in that it explicitly targeted at standardizing the use of Science Citation Index (SCI).

Topic trends

Fig. 2 shows the topic trends of the development of china's S&TEP. The size of the circles indicates the number of policies corresponding to each topic.

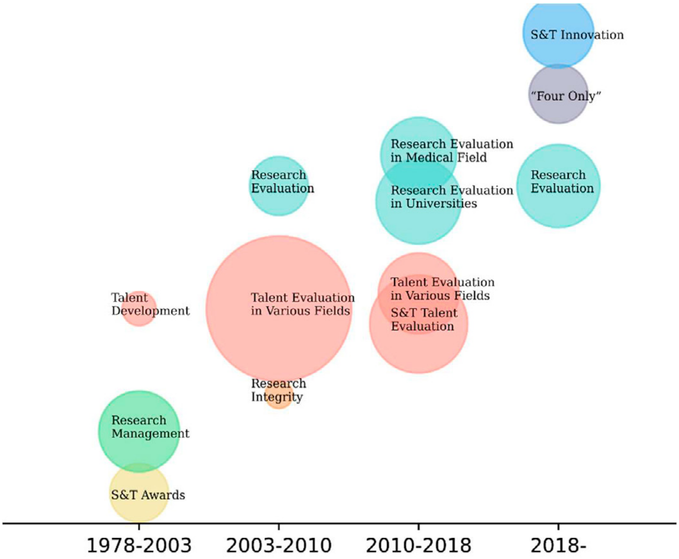


Figure 2. The topic trends of the development of china's S&T evaluation policies.

First, the figure shows that China's S&T evaluation policies involves various aspects such as S&T awards (A2), talent evaluation (A1, B3, C3, C4), research management and evaluation (A3, B2, C1, C2, D1), and research integrity (B1). Talent evaluation and research evaluation are two important topics evenly distributed in all stages, and involved fields are expanded in the development process. Secondly, combined with the keywords in each topic, it reveals that the policies in different periods are closely related to the social context in which they are located and the development of national S&T innovation. The fundamental driving force of the periodic transformation of S&TEP is the inability of the policies to adapt to the current situation of S&T development. After China's reform and opening up, S&T development is back on the agenda. Evaluation activities such as awards and titles are in urgent need of support from the evaluation system, so China's S&TEP has entered a period of exploration, achieving a transformation from non-existence to existence. The traditional peer-review method was mainly adopted at the stage. With the introduction of quantitative indexes such as SCI and CSSCI, quantitative evaluation methods began to be emphasized. However, imperfect evaluation system and the abuse of quantitative indexes leads to various problems, which prompting S&TEP to enter the stage of standardization. In this stage, evaluation policies were gradually standardized and the quality-oriented classification evaluation was proposed. Since the beginning of the 12th Five Year Plan, china formulated long-term plans for talents and education, which has greatly promoted S&TEP into the stage of innovative development. Innovation and originality become the primary evaluation criteria. Nowadays, China's modernization has put forward higher requirements for national science and technology innovation capacity, the proposal of breaking the *four only* and SCI-oriented system is regarded as the key point, and S&T evaluation activities are moving towards a new stage of comprehensively breaking the quantitative evaluation system and building a new evaluation system with Chinese characteristics.

In general, China's S&TEP has moved from singularity to diversification, from imperfection to standardization, from peer review-based to quantitative evaluation stage to the present stage of breaking quantitative evaluation and building a comprehensive system.

Public understanding of China's S&TEP

On 23, February 2020, the Ministry of Education and the Ministry of Science and Technology published and disseminated *Suggestions on the Correct Use of SCI as Research Evaluation Criterion and the Establishment of Proper Guidelines for Research Evaluation* (hereafter referred to as *Suggestions*), aiming to remove the distortions in the research evaluation system. These suggestions addressed the disproportionate, excessive and distorted use of SCI in research evaluation, offering ten suggestions and measures as guidelines for the use of SCI in academic evaluation. Since *the Suggestions* discouraged the current regulations on SCI papers and related indicators in the S&T evaluation, it has aroused intense public discussion. Numerous comments about this policy have emerged in various media. These data are an important basis for studying the public's response to this policy. We compare key points between policy comment from different sources. The results are shown in Table 6.

The commentaries in the newspaper are mainly written by reporters through interviews with scholars engaged in scientific research evaluation, or by invited staff members as guest commentators. Newspaper reports discussed the reasons and objectives of this policy and how China's scientific research evaluation should be reformed in the future. Firstly, the reason for problem of the current SCI-oriented evaluation system is that this indicator is directly linked to the evaluation of titles and resource allocation, thus leading many Chinese researchers to blindly pursue the number of papers instead of sinking their hearts to research. Second, some scholars analyze that the purpose of *the Suggestions* issued by the state is to break the situation of blindly imitating the SCI journal evaluation system and to construct evaluation methods, indicators and tools in line with China's value orientation. One of the important points is to create domestic academic journals with high quality and abstract more outstanding research achievements published in them. Third, scholars also discuss direction of reform in the future from various perspectives. For example, in the evaluation of research achievements, more attention should be paid to the construction of multiple expressions of research results and corresponding evaluation standards; In terms of talent evaluation, it is an effective way to implement masterpiece system and classification evaluation system to replace the old way of simply counting SCI papers.

Table 6. Comparison of key points of policy comment from different sources.

Content		*Newspaper*	*Scientific papers*	*Blog*	*Foreign media*
Similarities	Background	1. The singularity of evaluation indexes; 2. The phenomenon of researchers' eagerness to make profits			
	Purpose	1. Create and improve domestic academic journals; 2. Constructing evaluation methods, tools in line with Chinese values			
	Suggestions	1. The establishment of a diversified assessment framework; 2. Maintaining the quality and integrity of peer-review			
Differences	Purpose	Doesn't mean abolishing the SCI	Improving domestic academic journals; Constructing Chinese evaluation methods	Abolishing SCI in evaluation	Abolishing SCI in evaluation
	Suggestions	-	How to do with journals, research evaluation, especially in universities	Originality-oriented evaluation	-

The content discussed in Chinese scientific papers is quite different from the report. It is mainly about how to implement China's research evaluation activities in the future after the policy is released. Specifically, the issues discussed include the development, quality improvement and evaluation of Chinese academic journals, how to adjust the research evaluation system, and the evaluation of scientific research in universities and various disciplines.

For social media, we analysed 48 blogs and comment data after each blog. The topics discussed by scholars are roughly the same as newspapers and academic papers. But there is a different point of view. One of the points is that SCI did nothing wrong. Many scholars point out that SCI has made a great contribution to China's research evaluation system for a long time in the past. Its emergence has played an important role in breaking the circle culture formed by the old forces in the academic world. The problem with China's S&T evaluation is not the SCI, but the simplification of evaluation. It is worth mentioning that there are some scholars whose opinions about this policy are quite different from the official media's interpretation. They believe that the fundamental purpose of this policy is to abolish SCI in evaluation rather than just to weaken the importance of SCI in evaluation activities. Considering a series of policies recently released, one of the measures after the elimination of "only papers" is to attach importance to domestic scientific journals with international influence. In addition, some blogs analyse that the essence of moving away SCI-oriented evaluation system is to break the privileges and discrimination in S&T evaluation. An original and scientific comprehensive evaluation system should be established. That is, scientific research should be aimed at solving practical problems, not publishing papers. Scientific research results must be innovative and reflect service contributions.

Foreign media have paid great attention to China's S&T evaluation system. However, they analyse this policy from a more international perspective. On the one hand, scholars believe that this policy reflects China's new attitude towards SCI. The purpose of China's move is to abolish China's scientific research evaluation system that uses SCI as a core evaluation indicator, thereby helping the development of the domestic academic journal industry. On the other hand, scholars consider that this move may affect the participation of Chinese scholars in international academic cooperation, causing China to derail the development of international academics.

Overall, for all types of commentary texts, official media such as newspapers and academic papers interpret the research evaluation policy; while an important part of scholars' discussions about the research evaluation policy in social media and foreign media is the real purpose of this policy. To a certain extent, this difference reflects the policy tendency contained in the scientific research policy under China's authoritarian state.

Discussion and conclusions

Based on the text of S&TEP, this research explains the evolution and the focus of the China's S&TEP in each stage by using the method of LDA theme model. The results indicate that, China's S&T evaluation system has roughly gone through four stages, the trend of which is shift from publication number to publication quality. At present, China's scientific research evaluation system has entered a stage of deepening reform, with a trend of abandoning quantitative evaluation and constructing comprehensive evaluation system. Next, we take the policy "moving away only SCI evaluation system" in early 2020 as an example. From the four sources of policy comment texts, we can see that scholars have proposed the background, reason, purpose and future development of the S&T evaluation system of this policy from different perspectives. What's more, the text of China's S&TEP has the implications beyond the text itself, but they are easy to reach common acceptance among Chinese scholars.

The results of this systematic analysis are instructive for our understanding of China's S&T evaluation activities in the past, present, and future. The development of China's S&TEP is a long-term and continuous process of improvement. The current and future China's S&T evaluation system will focus on building a comprehensive and flexible evaluation system that remain focusing on originality and research quality. Promoting the construction of a scientific, fair and reasonable peer review system may be one of effective solutions after breaking evaluation framework with only quantitative indicators. Besides, it is also worth considering how to keep up with international academic exchanges while build an evaluation system with Chinese characteristics.

This study has two main contributions. First, this study provides a new division of the stages of China's S&TEP and reveals the evolutionary trends of research evaluation from the perspective of policy themes, providing new ways to understand the reform of China's S&T evaluation policies. On the other hand, we innovatively use policy comments data to broaden the data sources for S&TEP analysis. In our work, we only conducted qualitative discussion on policy review data. Since the review data is of great significant, we will consider using text mining technology to further mine the information contained in the review data in future work.

References

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations, Berlin, Heidelberg.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5), 993-1022. doi:10.1162/jmlr.2003.3.4-5.993

Cao C. (2004). China's Scientific Elite. *Taylor and Francis*.

Cao, C. (2008). China's brain drain at the high end: Why government policies have failed to attract first-rate academics to return. *Asian Population Studies*, 4(3), 331–345.

Chen, Y., & Zhang, Z. (2020). Opinions on New Science and Technology Evaluation Methods. *Journal of the China Society for Scientific and Technical Information*, 39(8), 796-805.

Du, B., &Wang, X. (2020). Analysis on Evolution Logic and Characteristics of Science and Technology Evaluation in China--the Text Analysis of Seventy Years of Relevant Researches. *Science & Technology Progress and Policy*, (07),120-129.

Gou, G., & Qi, Z. (2020). Review and Prospect of Academic Evaluation Policy in the Past 70 Years since the Founding of the People's Republic of China. *University Education Science,* (01),80-88.

Hofmann, T. (1999). *Probabilistic latent semantic indexing*.

Li, Q. (2020). The End of Publish or Perish? China's New Policy on Research Evaluation. https://www.mpiwg-berlin.mpg.de/observations/1/end-publish-or-perish-chinas-new-policy-research-evaluation

Li, Y., & Liu J. (2018). The Practice and Prospect of the Construction of Scientific and Technological Talent Team in China Since the Reform and Opening-up. *Human Resources Development of China*, 35(11), 30-43.

Liu, X. (2020). Structural changes and economic growth in China over the past 40 years of reform and opening-up. *China Political Economy*.

Mallapaty, S. (2020). China bans cash rewards for publishing papers. *Nature*, 579(7797), 18-18.

Miyata, Y., Ishita, E., Yang, F., Yamamoto, M., Iwase, A., & Kurata, K. (2020). Knowledge structure transition in library and information science: topic modeling and visualization. *Scientometrics*, 125(1), 665-687. doi:10.1007/s11192-020-03657-5.

Qian, J., Yuan, Z., Li, J., & Zhu, H. (2020). Science Citation Index (SCI) and scientific evaluation system in China. *Humanities and Social Sciences Communications*, 7(1), 108. http://doi.org/10.1057/s41599-020-00604-w

Quan, W., Chen, B., & Shu, F. (2017). Publish or impoverish an investigation of the monetary reward system of science in China (1999-2016). *Aslib Journal of Information Management*, 69(5), 486-502. doi:10.1108/ajim-01-2017-0014

Shao, J., & Shen, H. (2012). Research assessment and monetary rewards: the overemphasized impact factor in china. *Research Evaluation*, 21(3), 199-203.

Wang, Y. (2018). Historical Transformation of Chinese Science and Technology—National Science and Technology Conference of 1978 Revisited. *Bulletin of Chinese Academy of Sciences*, 33(04):351-361.

Xu, F., Gong, X. & Li X. (2018). 40 Years Reform and Development of Research Evaluation in China: Based on Case Studies onNSFCPeer Review and CAS Research Institutes' Comprehensive Evaluation. *Science of Science and Management of S.& T.*, (12),17-27.

Ye, J. (2020). The Keys,Roots,and Solutions toSCI Supremacy. *Journal of the China Society for Scientific and Technical Information*, 39(8), 787-795.

Zweig, D. (1997). To return or not to return? Politics vs economics in China's brain drain. *Studies in Comparative International Development*, 32(1), 92–125.

Zweig, D. (2006). Competing for talent: China's strategies to reverse the brain drain. *International Labour Review*, 145(1–2), 65–90.

Zweig, D., & Wang, H. (2013). Can China bring back the best? The communist party organizes China's search for talent. *The China Quarterly*, 215, 590–615.

The Effects of Rhetorical Structure on Citing Behavior in Research Articles

Wen Lou[1], Zilong Su[1], Shaodan Zheng[1] and Jiangen He[2]

[1] *wlou@infor.ecnu.edu.cn, 51204407348@stu.ecnu.edu.cn, daneezsd@163.com*
East China Normal University, Department of information management, Shanghai (China)

[2] *jiangen@utk.edu*
School of Information Sciences, The University of Tennessee, Knoxville, USA

Abstract

Aiming to explore the distributions of reference location in rhetorical sections of research articles and examine the relationship between the selection of rhetorical structure and the characteristics of references, we conducted a scheme with six types of rhetorical roles of literature review in research articles and applied to 4,931 JASIST research articles. The results showed that the pattern of selecting rhetorical structure of literature review in JASIST have evolved three stages from citing throughout to conventional literature review. Conducting a standard literature review section has become predominant for preparing scientific papers since early 21st century. We also found that the selection of rhetorical structure regarding literature review affect the disciplinary distribution of cited references. We will further investigate whether the rhetorical roles of distribution are related to the individuals' knowledge structure in selecting references.

Introduction

To facilitate scholarly communication in the context of globalization and international collaboration in science, conventional rhetorical structures are used by scientists in the field of their scientific interest (Kanoksilanpatham, 2005). Swales' approach to genre analysis has enhanced our understanding of rhetorical structure of the research articles (RA) that follow the conventional IMRaD (Introduction, Methods, Results, Discussion) macro-structure (Swales, 1990). Many studies have analyzed the rhetorical structure of individual sections of RAs in various disciplines, such as studies of computer science (Posteguillo, 1999), applied linguistics (Yang and Allison, 2004), biochemistry (Kanoksilanpatham, 2005), and comparative studies of multiple disciplines (Lin and Evans, 2012). In a microsense, rhetorical moves in each of the four main sections of RAs have been also well-studied (Bruce 2008; Peacock, 2012). As far as Literature Review section is concerned, it is less considered in the macro-structure in these analyses of RA rhetorical structure, though it is a common section and a necessary argumentative step in RAs. The function of Literature Review is also implemented as rhetorical moves in Introduction section. According to the Swales' CARS model (Swales, 1990), RA introductions contain three obligatory moves: (1) establishing a territory; (2) establishing a niche; and (3) occupying the niche. Move 1 and Move 2 of an RA introduction may fully implement the function of a Literature Review if there is no separate Literature Review section in the RA. The variation and effects of literature review in rhetorical structure of RAs largely remains an unexplored territory.

References cited in rhetorical sections have been found to have different characteristics. For example, the age of references varies by section, with older references in the Methods and more recent references in the Discussion (Bertin, 2016). Highly cited can be mainly selected for Methods purpose (Small, 2018), but also can be largely found in Introduction, Literature Review, and Discussion (Ding, Liu, Guo and Cronin, 2013; Thelwall, 2019). Hu, Chen and Liu (2013) found that citations are more likely to locate in the section of Introduction. Therefore, the selection of rhetorical structure for RA may affect the citing behavior of what kinds of reference would be cited in scientific writing.

Previous studies have found some factors that affect the author's citation behavior. For example, Milojević (2012) concluded that the collaboration level is an important factor related to citing behavior, especially in fields with extensive collaborative practices. A great deal of research has been done on citing behavior is about the reason for citing references, such as citing is guided by self-interest (Merton, 1957) or personal character (Kaplan, 1965). Lipetz (1965) characterized 28 and Duncan, Anderson, and McAleese (1981) concluded 26 different reasons for citing references. There are also many studies on the citation motivation. Garfield (1965) divided citation motivation into 15 categories, such as paying homage to pioneers, giving credit for related work, identifying methods, etc. Later lots of studies have listed various citation motivations from different perspectives. (Jurgens, Kumar, Hoover, McFarland & Jurafsky, 2018).

The selection of rhetorical structure determines how we organize and present a literature review in RAs in significant ways. For example, the distribution of cited references in a RA is greatly affected if the RA has an independent literature review section. Besides, since authors tend to cite different references across rhetorical sections, we believe the rhetorical structure regarding literature review may affect what references are cited and the rhetorical structure of scientific articles may also be a negligible factor for explaining citing behavior. Even the potential implications of rhetorical structure in citation studies, to our best knowledge, we have not found existing studies have examined the impact of the selection of rhetorical structure on citing behavior. In this article, we identify the evolving pattern of rhetorical structure regarding literature review by exploring the main sections for literature review in RAs. We also examine the relationship between the selection of rhetorical structure and the characteristics of references. We aim to answer the following research questions:

(1) How does the role of literature review evolve in rhetorical structures of RAs?
(2) How does the selection of rhetorical structures affect what kinds of references being cited in the literature review?

Research design

Data collection

We selected the Journal of the Association for Information Science and Technology (JASIST), a well-known peer review journal in the field of library and information science (LIS), as the data source. Full text and metadata of 5,623 articles in JASIST from 1950 to 2018 were downloaded in June 2019. We excluded several types of articles including brief communication, opinion paper, call for papers, and biographies. 4,931 articles in total were finally considered in our study. The metadata of all the references of 4,931 articles were downloaded as well in Web of Science, including the authors, journals, and publication years. Web of Science only provides JASIST information since 1985. So, 127,356 references since 1985 in total were collected and classified into LIS and Non-LIS categories based on the Journal Citation Report classification.

Coding process

First, to build up a coding scheme in terms of rhetorical structure, we randomly chose 100 articles (six to eight articles in every ten years by a random number generator) as coding samples. We manually read through all the articles and summarized the resemblances of how references distribute in an article based on the literal names of section headings. The resemblances eventually formed into six types of rhetorical structure regarding literature review. Type A and B both contain literature review sections. Type C and D are in line with IMRaD format. Second, we trained the coder to learn the six types among the coding samples.

And the coder passed a validation test via another 50 articles by an accuracy of 92%. Finally, the coder manually classified all the rest of articles.

A. Articles with specific sections entitled Literature Review, Related Work, etc.;
B. Articles with specific sections about literature review without entitled as Type A, but as headings with relative topics;
C. Articles that include literature review in the introduction section. A brief but standard literature review was placed in the end of or spread through introduction;
D. Articles that include literature review in the background section. Several citations instead of a standard literature review were placed in the beginning of introduction;
E. Articles without literature review but citations are spread throughout the full text;
F. Articles without any reference.

Findings

Overview

5.84% of total articles did not cite reference. They are mostly introduction of new concepts, techniques, and case studies. The most popular rhetorical structure in JASIST is surprisingly Type E, which exists the whole publishing life of JASIST. 45.79% of articles have cited reference all over the article. Type E dominates all indicators in Table 1, except the number of total reference and per article. The conventional structure of literature review, Type A, is secondarily most applied in JASIST articles. It showed appearances in most of the years. We observe that literature review in the introduction and background sections can be fairly acceptable in scientific writing. They yielded to almost 25% of all articles. Interestingly, Type B is the least the choice among JASIST authors, yet it occupied the largest number of references per article. It indicates the topical wise of literature review can possess a large proportion of references in individual articles compare to other structure.

Table 1. Statistics of rhetorical roles of literature review

Type	# articles (N)	% total	# appearance year (A)	Ave # per appearance year (N/A)	# total reference (R)	Ref per article (R/N)
A	1,178	23.89%	51	23.10	43,760	37.15
B	83	1.68%	24	3.46	3,788	45.64
C	817	16.57%	48	17.02	31,760	38.87
D	307	6.23%	34	9.03	12,788	41.65
E	2,258	45.79%	69	32.72	35,260	15.62
F	288	5.84%	34	8.47	0	0.00
Total	4,931	100%	69	71.46	127,356	25.83

Note. Appearance year is a year when a type of rhetorical roles appeared.

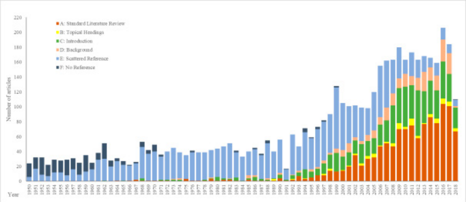
The distributions of rhetorical roles of literature review in research articles

The annual and proportional distributions certainly demonstrate couples of patterns with dynamic stages' shifts. We divide the total phrase of nearly 70 years into three stages.

Stage 1: Format-less stage (early time to the 1980s). Before scientific articles were published with a globally accepted format, citing behavior was even impossibly measurable and predictable. Although the IMRaD started to be adopted in science in the 1950s, the JASIST authors seemed to separate the roles of writing standard and citing preferences in this stage.

Stage 2: Formatting stage (the 1980s to the early 21st century). Along with the widely accepted standard for preparation of scientific articles (Day & Gastel, 2006), updated APA style (the only format accepted for JASIST all the time) version by version, and the rapidly increasing number of publications, the diversity of citing preferences in JASIST started to make a difference in this stage. Even though the scattered citing style (D) remained predominant, the emergence of other types of rhetorical roles began blossoming.

Stage 3: Standardized stage (the early 21st century till now). Conducting literature review sections in academic writing overthrew the domination of various cohabit citing behaviors. Citing reference in the conventional sections has become a trend in scientific publishing. Moreover, a norm of citing preferences in the specific literature review section appeared to be fundamental to publish in JASIST.



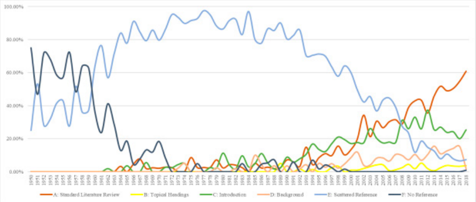


Figure 1. The annual distribution of rhetorical roles of literature review (left)
Figure 2. The annually proportional distribution of rhetorical roles of literature review (right)

Figure 3 compares the overtime changes of articles with literature review sections (Types A, B, C, D) and articles without literature review sections (Type E). Obviously, the proportion of articles with literature review sections has been increasing continuously. Although Type E dominated for the most of the time, especially before 2001, the turning point occurred in the early 1980s. This is when the American national standard for the preparation of scientific papers for written or oral presentation (ANSI Z39.16-1972) was published in 1972 and again 1979 (Day and Gastel, 2006). After 2001, the proportion of articles with literature review sections has begun to be higher than that of articles without literature review sections. The latter was below 50% for the first time. Presumably, the community of library and information science attracted and required researchers to submit manuscripts according to standard writing manual. Therefore, literature review has become an indispensable part in terms of scholarly communication standard since 2001. The growth rate increased from 2001 to 2010 and slowed down from 2010 to 2018 (a current peak has appeared). This slightly stability may result from the words limitation in the recent submitting instruction from Wiley.

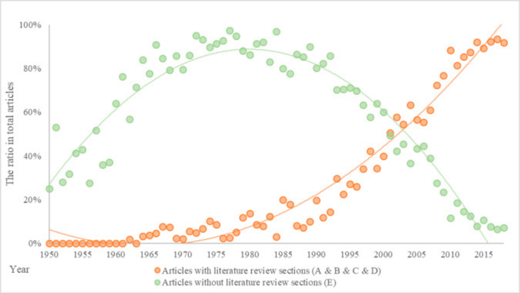
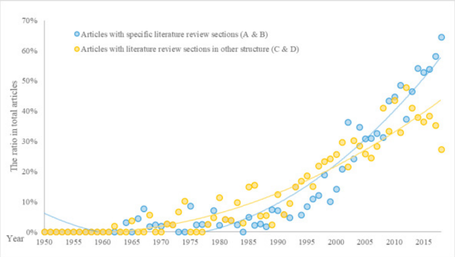


Figure 3. Comparison of articles with and without literature review sections

Note. The ratio is calculated from the total number of different types of rhetorical roles in one category dividing the total number of articles. Hereinafter as Figure 4 and 5.

Type A, B, C, and D all represent articles which contain literature reviews to some degree. Figure 4 shows the comparison between inner groups. The two trends are similar (Spearman correlation test result is 0.832, P<0.01) and generally show an upward trend. However, the slopes of them illustrate that the conventional literature reviews have drawn more attention than the other, especially after the early 21st century. The growth rate of those articles with specific literature review sections increased even more since 2012 with a current peak of 64.55% in 2018. This interests us. JASIST has started to limit word count currently. We presume the new policy would have affected the writing pattern regarding the selection of rhetorical structure in terms of literature review as literature review is normally considered as a word-demanding section. The results failed our assumption.



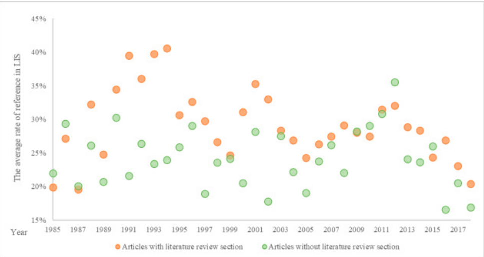


Figure 4. Comparison of articles with specific literature review sections and within other parts of manuscripts (Left)
Figure 5. Comparison of reference in articles with and without literature review sections (right)

The relationship between rhetorical structures and references

The analyses above have given the distributions of rhetorical structures from the perspective of publications. We collected the references of all the publications to examine the distributions of the relationship between rhetorical structure and references. We composed a comparison in Figure 5 between Type A and Type E, respectively representing articles with and without literature review sections. We assume that articles with literature review section would contain more LIS references (references were published in LIS journals) than those without literature review section. Mann - Whitney U test was conducted to examine whether articles with literature review section would have relationship with the results of these articles have cited. The results show the significant difference between the articles with literature review sections and the ones without such sections ($U = 274$, $z = -3.738$, $P = 0.000 < 0.005$). Apparently, articles with literature review section have distributed on the higher rate than the other in Figure 5 (logistic regression result shows the difference is 5.83 times). This suggests the conventional literature review clusters inner-disciplinary topics rather comprehensive topics.

Conclusions and future research

In this article, we made a preliminary statistics and analysis on the distribution of the literature review in rhetorical sections. The pattern of selecting rhetorical structure of literature review in JASIST has been three stages from citing throughout to conventional literature review. Conducting a standard literature review section has become predominant for preparing scientific papers since early 21st century. Our preliminary results also show that the rhetorical structure may have effects on the disciplines where cited references are from. Future studies will consider how the selection of rhetorical structure affects selecting the types of cited references, and how these citations in turn impact on the knowledge structure of the essay. Furthermore, we will examine the selection preferences of rhetorical structure from the authors' perspective and disciplinary differences.

Acknowledgements

This research was supported by China Postdoctoral Science Foundation (2019T12035).

References

Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016). The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology*, 67(1), 164-177.

Bruce, I. (2008). 'Cognitive genre structures in methods sections of research articles: A corpus study,' *Journal of English for Academic Purposes* 7/1: 38–54.

Day, R.A. & Gastel, B. (2006). *How to write and publish a scientific paper (6th ed)*. Greenwood Press, Westport.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis, *Journal of Informetrics*, 7(3),583-592,

Duncan, E.B., Anderson, F.D., & McAleese, R. (1981). Qualified citation indexing: Its relevance to educational technology. *Proceedings of the First Symposium on Information Retrieval in Educational Technology*, pp, 70–79.

Garfield, E. (1964). Can Citation Indexing Be Automated? *Symposyum on Statistical Assoc Methods for Mechanized Documentation*, 84-90.

Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics*, 7(4), 887-896.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D. & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6,391-406.

Kanoksilapatham, B. (2005). Rhetorical structure of biochemistry research articles. *English for specific purposes*, 24(3), 269-292.

Kaplan, N. (1965). The norms of citation behavior: prolegomena to the footnote. *Journal of the Association for Information science and Technology*, 16(3), 179-184.

Lin, L. and S. Evans. 2012. 'Structural patterns in empirical research articles: A cross disciplinary study,' *English for Specific Purposes* 31/3: 150–60.

Lipetz, B. A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *Journal of the Association for Information Science and Technology*, 16(2), 81-90.

Merton, R. K. (1957). Priorities in scientific discovery: a chapter in the sociology of science. *American Sociological Review*, 22(6), 635-659.

Milojević, S. (2012). How are academic age, productivity and collaboration related to citing behavior of researchers? *PloS One*, 7(11), e49176.

Peacock, M. 2002. 'Communicative moves in the discussion section of research articles,' *System* 30/4: 479–97.

Posteguillo, S. 1999. 'The schematic structure of computer science research articles,' *English for Specific Purposes* 18/2: 139–60.

Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *Journal of Informetrics*, 12(2), 461-480.

Swales, J. M. 1990. Genre Analysis: English in Academic and Research Settings. Cambridge University Press.

Thelwall, M. (2019). Should citations be counted separately from each originating section? *Journal of Informetrics*, 13(2), 658-678.

Yang, R. and D. Allison. 2004. 'Research articles in applied linguistics: Studies from a functional perspective,' *English for Specific Purposes* 23/3: 264–79.

Wu, J. (2011). Improving the writing of research papers: IMRaD and beyond. *Landscape Ecology*, 26(10), 1345-1349.

No Metrics for Postdocs: Precarious Labour in Science Policy

Lai Ma[1]

[1] *lai.ma@ucd.ie*
School of Information and Communication Studies, University College Dublin, Belfield, Dublin 4, Ireland

Abstract

In recent years, the pressure of producing impacts such as the creation of intellectual property and other commercialisation activities ('knowledge transfer') has increasingly dominated the discourse of research institutions and universities. Research projects can be comparable to 'gigs' when they employ postdocs on precarious fixed-term contracts. However, there seems to be little consideration in research and science policy about the career development of postdocs beyond funded projects and there seems to be no metrics about the contributions of postdocs to knowledge production, nor data about 'brain drain' as a result of precarious contracts. Using in-depth, semi-structured interviews with postdocs, PIs, and support staff, this study aims to understand the perceived roles of postdocs as a career stage and the perceived success factors that help them transitioning from precarious contracts to long-term academic/research positions. The work-in-progress paper will discuss some preliminary findings including the meanings and contexts of postdoc, as well as the problems and issues of precarious, fixed-term contracts in relation to publication and knowledge production. This paper also calls for comprehensive data collection and analysis about the contributions by postdoctoral researchers and the potential loss of knowledge as a result of the precariousness of academic career.

Introduction

For decades, scholarly works and scientific research have been driven by a reward system that primarily recognises research funding, research metrics and 'mobility' or internationalisation (Ackers, 2008) as key performance indicators. The reward mechanisms have lent powers to governmental bodies and funding agencies who set priority areas and allocate funding accordingly. The pressure of producing impacts such as the creation of intellectual property and other commercialisation activities ('knowledge transfer') has increasingly dominated the discourse of research institutions and universities. Research projects are comparable to 'gigs' when they employ a large number of gig workers (PhD students, Postdocs) whose career paths are insecure and unclear. However, there seems to be little consideration in science and research policy about the career development of postdocs beyond funded projects. There is also no metrics about the loss of precarious labour—and the many contributions they would and could have made to science and society.

The recent OECD report (2020), *Resourcing Higher Education*, has highlighted the harmful consequences of extensive casualisation of academic staff, for example, low retention of researchers, teachers, and students and lower quality of teaching and learning. Kwiek (2019) has examined stratifications in academic performance and power, pointing to the need of predictable career advancement for early career researchers. Flynn (2020) has described the disheartening experiences of being precarious with no clear paths to obtain a research grant or land a permanent position years after earning her PhD. The casualisation of work also limits the participation of the 'gig workers' in university and research governance.

The development and challenges faced by early-career researchers, especially postdocs, have been studied in terms of regimes of valuation (Fochler, Felt and Müller; 2016), symbolic violence (Roumbanis, 2019), practices of appraisal devices (Nästesjö, 2020) and so forth. Most recently, a survey of postdoc conducted by *Nature* (Woolston, 2020) reveals 'great distress' experienced by postdoctoral researchers worldwide. Notably, Milojević, Radicchi and Walsh

(2018) have found dramatic shortening of careers of scientists from 35 years in the 1960s to only 5 years in the 2010s, some of whom as supporting authors only in their entire career. There is, however, a lack of metrics about the (loss of) productivity and performance of postdocs in this competitive research environments due to the limited number of academic and research positions.

This exploratory study aims to understand postdocs as a career stage and the perceived success factors that help transitioning from precarious contracts to long-term academic/research positions, as well as career beyond academia. This work-in-progress paper will discuss some preliminary findings of semi-structured interviews with postdocs and support staff in research institutions, focusing on the relationship between precariousness and knowledge production. It is the goal of this study to establish the need for future empirical and quantitative studies pertaining to the contributions of postdoctoral researchers in knowledge production.

Method

The first phase of the study was conducted between July and September 2020. Twenty in-depth, semi-structured interviews were conducted with postdoctoral researchers and research support staff in universities and research centres in Ireland. The list of potential participants was collated by searching university websites using the title, 'postdoc' and similar terms[1]. It is worth noting that, however, the contact information of postdoctoral researchers is not necessarily listed by research centres and universities. As a result, most respondents were receipents of the Irish Research Council (IRC) Postdoctoral Fellowship, partly because their affiliation and contact information have been made publicly accessible. The potential participants were invited to participate in the study by individual emails. An information sheet and an informed consent form were sent to respondents prior to the interviews.

During the interviews, the postdoctoral reseachers were invited to talk about their research career and their plans for the future, their experience as a postdoctoral researcher and the pros and cons of postdoc as a career stage, while the support staff were asked to comment on the needs of the postdoctoral resaerchers, including support for grant applications and career development. The interviews were conducted using Zoom or Skype and are approximately 35-70 minutes in length. They were transcribed fully first by otter.ai and then checked and corrected manually. The transcription is anonymized and coded based on emerging themes. The data collection will continue in early 2021 until saturation has been reached.

Preliminary Findings

In this section, emerging themes will be discussed based on the coding and analysis of the first phase of the study. These preliminary findings show the impact of precarious contracts on publications and related issues for postdoctoral researchers, highlighting the need for future quantitative studies to shed light on the scope and depth of the problems.

Postdoc as a Career Stage

Many senior faculty/academic staff in research institutions and universites have not been a postdoc themselves, for it was not considered as a bridge between doctoral studies and

[1] There is not a clear definition of 'postdoc'. Doctoral researchers can be working on funded projects with various levels of freedom in pursuing research interests of their own, while some— most respondents in this study—are awarded with fellowship to engage in individual projects. However, it should be noted that many individuals with a PhD can be working as research assistants, research fellows or under other titles in a research institution or university in Ireland. It is unclear as to whether they should be considered as 'postdoc' because they are not classified as academic staff.

permanent (or tenured-track) positions decades ago. Today, despite the lack of data about the percentage of faculty/academic staff with or without postdoc experiences, it is generally understood that newly graduated PhDs would be in at least one and indeed often multiple postdoctoral positions. Many respondents in this study reported that it is not uncommon to have 3-5 postdoctoral contracts before landing a permenant position, or before one decides to quit academia altogether.

Is postdoc a necessary a career stage? Most respondents said yes. However, the reasons demonstrate the complexity of 'surviving' in a very competitive environment—that it is considered impossible for anyone to land a position without working as a postdoc for several years, that it needs the time period as a postdoc to increase the number of publications for applying for grants and/or permanent positions, that it is a career stage where one can develop networks considering the preferences for internationalisation and mobility in research career (see Archer 2008).

For respondents who have been awarded an individual fellowship (e.g., IRC, Marie-Curie), the postdoctoral position affords them time and space to prepare publications based on their PhD studies or embark on a new project. These respondents were conscientious of the privilege of the fellowship, for they do not have obligations to undertake teaching or administrative responsibilities in the host institutions. In other words, they can devote all of their working hours to pursue their own research projects and they are also avail of research budget for buying materials and traveling to archives and conferences. Many stated that the time offered by the fellowship has been essential for them to develop as an independent researcher/scholar, partly due to the projectification of doctoral training (see Torka, 2018). For respondents who are working on funded projects, the postdoctoral position affords them to acquire and develop skills before applying for grants as a principle investigator (PI). The respondents mentioned project management and supervision as essential skills in their career development—whether they plan to pursue a career in research institutions or industry. Many reported that they are also developing expertise in methodologies and techniques as a postdoc.

Timeframe

While the respondents articulated the many benefits a postdoctoral position can offer, all commented that the duration of postdoctoral contracts are often too short. Most agreed that a three-year contract is about optimal while shorter contracts tend to create stress and sometimes mental health issues due to the following reasons: first, the lack of job security means that the postdocs would be looking for the next position and working on applications from Day 1 which constitutes a substantial workload in addition to the 'day job'. Second, their workload can be compounded by the pressure to publish from their PhD work or previous project(s)—with the assumptions that publications will eventually lead them to a permenant/stable position. Third, a new contract usually requires relocation, meaning the lack of support by family and friends. Indeed, as many have reported, an academic career often entails delaying family and personal plans due to the frequent relocation in different cities/counries at the postdoctoral stage, which usually lasts 5-6 years and sometimes longer[2].

The respondents reported that the precariousness of the postdoctoral positions has negative impacts on knowledge production as they find it difficult to finish writing from previous projects, and/or they sometimes cannot finish a study due to the time constraint of the contracts. There are also cases where one does not have an affiliation and hence loses access to materials

[2] The estimate of 5-6 years is based on anecdotes as there is a lack of official data.

and support provided by academic libraries and other support units. Many lamented that the short and rigid timeframes simply do not work for the nature of scientific research and scholarly inquiries, which can have unexpected delays or take unexpected turns.

The Pseudo-Employee Status

The precarious, fixed-term contracts also affect the employee status of postdocs. Since postdocs are mostly funded by research councils or other funding agencies, there are no commitments or obligations for the employers to retain the postdocs after the fixed-term contracts. While some universities do provide career development support for postdocs, the options can be minimal compared to those provided for permanent staff. Some respondents reported that they do not have a sense of belonging to the university—which may be better described as 'host institution'—even though they are supposedly 'employees'. While some are affiliated with research hubs and communities, some are totally isolated. For instance, the respondents reported that they do not attend staff meeting and/or decision-making processes in their host institution and some felt that they were treated as 'second-class citizens'. During the intial stage of data collection of this study, it is also clear that postdocs are not necessarily listed as academic staff or faculty on university websites, meaning that often these postdocs cannot be easily found by a Google search. Some respondents maintain their own web presence by hosting a website themselves, or by using third-party services such as academia.edu.

Due to the timeframe of their contracts, the respondents' involvement in university life is usually limited to their research group, if any. Some respondents were aware of research staff associations, but commented that their activities can be sporadic, and may be discontinued, when active members leave. At the same time, they are also most concerned with their career and spare little time on social activities. Postdoctoral researchers can be seen as a group of 'gig workers' who receive limited benefits and have no say on university and research governance, while their contributions to research and knowledge production have been under-documented and understudied.

Summary and Future Studies

This study aims to understand the postdoctoral experiences in the context of science and research policy—Is postdoc a necessary career stage? What are the benefits and challenges? What are the implications for knowledge production and the future of scientific research and scholarly inquires? The respondents reported that their postdoctoral experiences have been useful for concentrating on a project, learning new skills, and most importantly, developing into an independent researcher/scholar. Postdoc as a career stage, however, can be attributed to the competitive market, that is, the lack of academic positions, for one can argue that PhD graduates can develop their projects and skills in a permenant/tenured-track position without going through precarious contracts. There are many epistemic and labour issues pertaining to the fixed-term, precarious contracts of postdocs, including the loss of knowledge when one switches from one contract to another, or when one leaves the academic/research career altogether. As of now, however, there is no data or metrics recording the contributions to publications by postdoctoral researchers, nor data about the potential loss of knowledge—brain drain—due to precarious contracts. As Stephan (2013) has aptly pointed out, 'the low price of postdocs hides the true cost of postdocs to society' (p. 245). While warnings have been raised about "postdocs in crisis" during the pandemic (Nature Editorial, 2020), data and analysis about the publication patterns of postdocs can unearth issues about the consequences of precarious contracts on scholarship and scientific progress.

References

Ackers, L. (2008). Internationalisation, mobility and metrics: A new form of indirect discrimination? *Minerva*, *46*(4), 411–435. https://doi.org/10.1007/s11024-008-9110-2

Flynn, D. (2020). On being precarious. *Irish University Review*, *50*(1), 6.

Fochler, M., Felt, U., & Müller, R. (2016). Unsustainable Growth, Hyper-Competition, and Worth in Life Science Research: Narrowing Evaluative Repertoires in Doctoral and Postdoctoral Scientists' Work and Lives. *Minerva, 54*(2), 175-200.

Kwiek, M. (2019). *Changing European Academics: A Comparative Study of Social Stratification, Work Patterns and Research Productivity.* London and New York: Routledge

Milojevic, S., Radicchi, F., & Walsh, J. P. (2018). Changing demographics of scientific careers: The rise of the temporary workforce. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(50), 12616–12623. https://doi.org/10.1073/pnas.1800478115

Nature Editorial (2020). Postdoc in crisis: Science risks losing the next generation. *Nature*, *585*, 160. doi: https://doi.org/10.1038/d41586-020-02541-9

Nästesjö, J. (2020). Navigating Uncertainty: Early Career Academics and Practices of Appraisal Devices. *Minerva*, 1–23. https://doi.org/10.1007/s11024-020-09425-2

OECD (2020). *Resourcing Higher Education: Challenges, Choices and Consequences*, Higher Education, OECD Publishing, Paris, *https://doi.org/10.1787/735e1f44-en*

Roumbanis, L. (2019). Symbolic Violence in Academic Life: A Study on How Junior Scholars are Educated in the Art of Getting Funded. *Minerva*, *57*(2), 197–218. https://doi.org/10.1007/s11024-018-9364-2

Stephan, P. (2013). How to Exploit Postdocs. *BioScience*, *63*(4), 245–246. https://doi.org/10.1525/bio.2013.63.4.2

Torka, M. (2018). Projectification of Doctoral Training? How Research Fields Respond to a New Funding Regime. *Minerva*, *56*(1), 59–83. https://doi.org/10.1007/s11024-018-9342-8

Woolston, C. (2020). Postdoc survey reveals disenchantment with working life. *Nature*, *587*, 505-508. doi: https://doi.org/10.1038/d41586-020-03191-7

Studying researchers' institutional mobility: bibliometric evidence on academic inbreeding and internationalization

Vít Macháček[1], Martin Srholec[2], Márcia R. Ferreira[3], Nicolas Robinson-Garcia[4] and Rodrigo Costas[5]

[1]*vit.machacek@cerge-ei.cz*
Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague (Czech Republic)

[2]*martin.srholec@cerge-ei.cz*
CERGE-EI, a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences, Prague (Czech Republic)

[3]*ferreira@csh.ac.at*
Complexity Science Hub Vienna, Vienna (Austria)

[4]*elrobinster@gmail.com*
Information and Media department, University of Granada, Granada (Spain)

[5]*rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Leiden University, Leiden (the Netherlands)
DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, Stellenbosch (South Africa)

Abstract:

We propose institutional mobility indicators based on researchers' mobility flows in 22 major fields of science across 1,130 Leiden Ranking institutions from 64 countries. We base our indicators on data from the Dimensions and GRID databases. We use researchers' first and last affiliations to estimate the extent authors have moved across institutions as well as universities' degree of internationalization. For each institution, we quantify the shares of researchers with the same affiliation (insiders), those who came from another institution within the country (domestic outsiders) and those coming from a different country (foreign outsiders). Institutions in Central, Eastern and Southern Europe have the highest share of insiders, whereas institutions in Northern America and Western and Northern Europe have a higher share of foreign outsiders. Foreign outsiders are most common in small and wealthy countries. No disciplinary differences are observed, as captured by the field classification scheme of Dimensions.

Introduction

There is an increasing need to account for how institutions internationalize (Hoekman, 2012; Robinson-Garcia et al., 2019). Evidence suggests that cross-national mobility flows are increasing (Sugimoto et al., 2017) as is the strengthening of global collaborative research networks (Wagner and Leydesdorff, 2005). Research institutions foster internationalization as they compete to attract and retain foreign talent in a global market (Hazelkorn, 2011; Seeber et al., 2016), with foreign born researchers accounting for 42% of life sciences postdoctoral research in Europe, and 57% of these coming from outside the European Union (Moguérou and Di Pietrogiacomo, 2008). As a result, there is a policy interest in monitoring and understanding the process of academic mobility and internationalization of universities (OECD 2008, European Commission 2012, Jacob and Meek 2013, Sugimoto et al. 2017).

The lack of global and harmonized datasets has been a persistent challenge in developing global indicators of mobility (Sugimoto et al., 2016; Welch et al., 2018). This is especially problematic at the institutional level, due to the difficulties in obtaining fine-grained mobility data and lack of scalable methods. Recently, bibliometric databases have significantly improved the consistency of publication metadata, particularly author-affiliation linkages, allowing the tracking of affiliations of individual researchers and opening up new opportunities for studying long-term mobility patterns at scale (Moed and Halevi 2014, Sugimoto et al. 2016). Previous work has been concerned exclusively with the quantification of movement of scholars across countries and disciplines (Aref et al., 2019; Robinson-Garcia et al. 2019). In this paper we attempt to develop mobility indicators for research institutions.

In this paper we deliver novel insights on the composition of the academic workforce of universities and national research systems worldwide. We base our results on the Dimensions database and the Global Research Identifier Database (GRID). We propose three indicators by which institutional workforces can be characterized based on researchers' first institution of affiliation. By comparing the institution to which researchers were affiliated in one of their first publications with the most recent one, we distinguish between:

i) *insiders*, defined as researchers who are currently affiliated to the same institution to which they were affiliated in one of their first publications;
ii) *domestic outsiders*, that is, those who were originally affiliated to a different institution within the same country from their current one; and,
iii) *foreign outsiders*, researchers who were originally affiliated to an institution located in a different country from their current one.

The paper is structured as follows. First, we review the existing literature on academic mobility and the related concept of inbreeding. Then we describe the mobility indicators with a description of how we built the dataset. Next, we present the empirical findings from an aggregate perspective and analyze the heterogeneity of universities with regards to researchers' institutional mobility between countries and across the world. We conclude by discussing implications, limitations of this approach and future research lines derived from this study.

Literature review

Academic mobility is widely perceived as beneficial to the scientific enterprise (Wagner & Jonkers, 2017; European Commission/EACEA/Eurydice, 2015). It is considered important to promote the dissemination of knowledge, acquiring experience in different research environments, career advancement, and for creating opportunities for collaboration (e.g., Sugimoto et al., 2017; Stephan & Levin, 2001; Wagner & Jonkers, 2017). Mobile researchers serve as important bridges between countries (Meyer, 2001), they reinforce existing or create new ties with other national and foreign institutions, thus generating collaborative research networks that span the globe. When researchers move, they bring with them knowledge and ideas that differ from natives, which are essential for knowledge recombination and interactive learning, which in turn may lead to innovation (Ganguli, 2015; Stephan & Levin, 2001).

The notion of academic mobility is closely related to the concept of internationalization (Wagner & Jonkers, 2017). The process of internationalization has become an essential part of universities' strategic planning as it involves the implementation of programs that support the periodic movement of researchers, building ties with top universities, and improving visibility. The increasing orientation towards internationalization is in part due to strategic considerations in the context of a global competition for talents (Seeber et al., 2016). Internationalization strategies are particularly used by universities in developed countries (Lepori et al., 2015), which are increasingly dependent on the movement of researchers as a way of maintaining their attractiveness and international reputation.

The lack of institutional internationalization - which translates into low workforce mobility in a given institution - is thought to be decisive in influencing academic productivity and improving researchers' performance (Franzoni et al., 2014; Horta et al., 2010; Horta, 2013) and citation impact (Ganguli, 2015; Sugimoto et al., 2017). A recent large-scale bibliometric study reported that mobile researchers (more than one affiliation to different countries) were more cited than non-mobile researchers (Sugimoto et al., 2017). A similar finding was report by Ganguli (2015) who found that Russian scientists who have migrated to the US have been more cited by US scientists than those who have not migrated. In the US context, Foreign-born faculty has also been found to publish more (Mamiseishvili & Rosser 2010) and publish more breakthroughs (Stephan & Levin, 2001) than their US-born counterparts. One explanation for lower levels of impact and productivity of immobile researchers is that this kind of faculty has less access to varied information and exchange dynamics which may explain their lower scientific productivity (Horta, 2013).

Other studies suggest that academic mobility may or may not have a negative effect on the researchers' careers. For example, Melin (2005) has shown that mobile scholars returning to Sweden after a postdoctoral period abroad do not benefit from their foreign experience. Cruz-Castro and Sanz-Menéndez (2010) have found that non-mobile careers are a strong predictor of the timing of rewards in the form of early permanent positions in Spain, while mobile careers in the US increases researchers' chances of obtaining tenure. This suggests that the relationship between mobility and scientific performance varies widely across national research systems (Cruz-Castro & Sanz-Menéndez, 2010).

A related stream of literature contrasts mobility with the practice of having PhD graduates employed by the university that also trained them. This is known as "academic inbreeding" (Yudkevich & Sivak, 2012; Caplow & McGee, 1958; Berelson, 1960) or "institutional inbreeding" (Horta, 2013). Like the lack of mobility and internationalization, inbreeding is seen as negative for both the institution and researchers and has also been an indication of poor institutional performance (Horta, 2013). For example, academic inbreeding has been associated with lower output (Horta et al., 2010), fewer articles published in peer-reviewed journals than non-inbred (Horta, 2013), and fewer foreign co-authors (Scellato et al., 2012), effectively slowing down the career development of scholars (Inanc & Tuncer, 2011). Low productivity levels of inbred scholars emphasize the need for mobility and calls for policies to curtail academic inbreeding (Horta, 2013).

Data and methodology

The university's reliance on insiders among its workforce producing scientific results, can be interpreted as the university tendency to academic inbreeding. Empirical evidence on publication activity of PhD students indicates that even though differences between disciplines matter, the first papers' affiliation can be used as a proxy for Ph.D granting institution of the individual researcher. For example, Lee (2000), Lariviére (2012) and Waaijer et al. (2016) found that PhD students publish their first article before graduation quite frequently in natural and health sciences, though less so in social sciences and humanities. Nevertheless, PhD students may acknowledge affiliation to their alma mater, even if the dissertation research is published only after leaving the university due to publication delays in peer-review journals. Macháček and Srholec (2020) directly tested this assumption in a random sample of 90 researchers affiliated to major Western and Central European universities derived from the Scopus citation database. They found out from publicly available sources, from which university the researcher graduated, and compared that with outcomes of the bibliometric approach outlined above. The conclusion on whether the researcher is currently based on her alma mater matched by 77% in biochemistry, genetics and molecular biology, 90% in physics and astronomy and 87% in social sciences. There was of 7 false positives and false negatives, which tend to offset each other, thus the impact in aggregated data is even more limited.

Database

We use the Dimensions database version from June 2019 that is available in the CWTS database system, including data on publications, affiliations, researchers as well as disciplines (Herzog et al., 2020), which in total contains data for more than 100 million documents. Dimensions has a similar coverage to Scopus (Thelwall 2018), at least for publications with a Digital Object Identifier (DOI), which are predominantly journal articles but also book chapters, conference proceedings and others.

The Dimensions database has its own author name disambiguation procedure. In general, author name disambiguation algorithms attempt to group citation records of the same author by finding some similarity among them or try to directly assign publications to the individual who wrote them (Caron and van Eck, 2014). The Dimensions author name disambiguation algorithm is a two-step procedure. First, it uses "affiliation data, co-authorship and citation patterns as well as subject area traits" (Hook et al., 2018, p. 8) to produce clusters of publications that belong to potential individuals. These clusters are then connected using Open Researcher and Contributor ID (ORCID) and DOIs. Each disambiguated author is then assigned a researcher ID by which is unique identifier, producing almost 20 million researchers. Researcher IDs have been successfully assigned to about 87% of publications-authors combinations (Hook et al. 2018).

However, author-disambiguation algorithms (including the ORCID) are prone to errors (Caron and van Eck 2014, Gurney et al. 2012, Levin et al. 2012), and there is a need to establish a balance between precision (i.e. *How many identified publications truly belong to the same researcher?*) and recall (*Are all researchers' publications correctly identified?*). The Dimensions algorithm favors precision over recall (Bode et al. 2018), which is particularly relevant for us, since mobility events can cause splitting researchers into multiple IDs, and the precision preference can lead to the under-estimation of mobility events. For example, the publications of a researcher are split in disconnected clusters across her affiliations because the algorithm may not be able to merge them together.

Dimensions combines publication metadata with its own database of harmonized research institutions – the Global Research Identifier Database (GRID). This allows us to easily track the movements of scholars from one institution to the next. Aside from the unique institution identifier, GRID provides detailed geographical information, such as coordinates and city, region, and country names and codes. Other attributes include the institution's year of establishment and its type, which is divided into eight categories, one of which identifies educational institutions and universities.

Dimensions also includes a field-classification scheme for publications based on Australian and New Zealand Standard Research Classification (Australian Bureau of Statistics, 2008), which is a three-level hierarchical system of categories. Because more detailed disaggregation leads to a limited number of observations in the sub-disciplines, we use the top-level of this classification (FoR-division) in this paper, which refers to 22 major disciplines across all fields of sciences.

Determining the mobility category of researchers

We used *2018* as the reference year to identify researchers affiliated with specific institutions. Furthermore, we only include researchers with sufficiently long publication histories (i.e., more than 6 years since their first publication). The first year of publication is used as a proxy for measuring academic age (Nane et al. 2017), preventing researchers with very short publication histories from driving the results. All researchers publishing their first paper in 2012 or later are excluded from the analysis.

Each researcher r can belong to multiple sets as researchers can be affiliated to multiple universities and publish in multiple disciplines. Consider researcher r is affiliated to institution u and publishing in discipline d in 2018. She has a set of publications $P \subset p_{p,t}$ where at least one $p_{p,t>6}$, in which p is the publication index and t is the index for the number of years since the first publication (t is calculated from the publication calendar year). To determine where the researcher r started her career we select a subset of her initial publications published in the first two calendar years of her publication history, where:

$$P^{start} = |p_{p,t} where\ t < 2|$$

Based on P^{start}, we derive our indicators – whether researcher r started her publication career at the same university ($r^{start,u}$) and country ($r^{start,c}$). $r^{start,u} = 1$ if r is affiliated to u in any publication in P^{start}, otherwise $r^{start,u} = 0$. Similarly $r^{start,c} = 1$ if r is affiliated to any institution in country c in any publication in P^{start}, otherwise $r^{start,c} = 0$.

This allows us to split researchers into three mutually exclusive categories. For each university u and discipline d we then report the share on the total number of researchers of:

1. Insiders: Researchers starting at the same university – $\{\mathrm{r}|r^{start,u} = 1\}$
2. Domestic outsiders: Researchers starting at another institution in the same country – $\{\mathrm{r}|r^{start,u} = 0 \text{ and } r^{start,c} = 1\}$
3. Foreign outsiders: Researchers starting abroad – $\{\mathrm{r}|r^{start,c} = 0\}$

Selecting research universities based on the Leiden Ranking

We only consider established research universities that are included in the Leiden Ranking. The GRID category of "educational institutions" is far too broad for this purpose. Therefore, we use the Leiden Ranking methodology to identify the most productive research universities worldwide (Waltman et al, 2012). This results in a fairly homogenous group of well-established universities with significant research activities. We use the Leiden Ranking 2020 data (see van Eck, 2020), which includes a total of 1,176 universities, only 7 of which we were unable to match with the initial set of GRID identifiers. While most universities have been created before World War II and only about a tenth of them in the mid-1970s or later, there are some universities in that were established in 1998 or later according to the GRID database. We removed those universities from our sample since younger, less established universities have had less time to employ their own graduates. Using a cut-off year other than 1998 has no effect on the main results and the final sample includes 1,130 universities.

We rely solely on author-affiliations linkages as reflected in researchers' publication record. That means that researchers will be labeled as *insider* regardless of the type of position they have at a given institution. For example, some students might have a first publication already in the undergraduate studies. Or some universities might not provide doctoral education in some disciplines. Unfortunately, there is no easy way to identify all this diversity of affiliation linkages, unless one connects the bibliometric records with administrative data, which is not feasible at the global scale.

Results and discussion

Overall, the share of insiders in total researchers differs markedly across the universities. On an average university nearly half of researchers hold the insider status, but this share ranges from less than a third in about one fifth of the universities to more than three quarters in a tenth of them; with some notable outliers at both ends of the spectrum. As can be expected, the flipside of employing initial outsiders also varies significantly with a clear predominance of locals over foreigners. Only about a tenth of the universities maintain more than a third of researchers who started publishing with affiliations abroad.

Figure 1 displays the share of insiders in universities worldwide. Each node denotes one university. The size of the node is proportional to the number of researchers in the respective university. The intensity of the colour reflects the share of insiders. Areas with the highest share of insiders are concentrated in the South and East regions, while low shares of insiders predominate in the North and West regions. Oceania, namely Australia and New Zealand, are the only exceptions. We also observe an overrepresentation of North American and European universities in contrast with other regions of the world such as Africa, Southeast Asia (except for China and Japan), and South America.

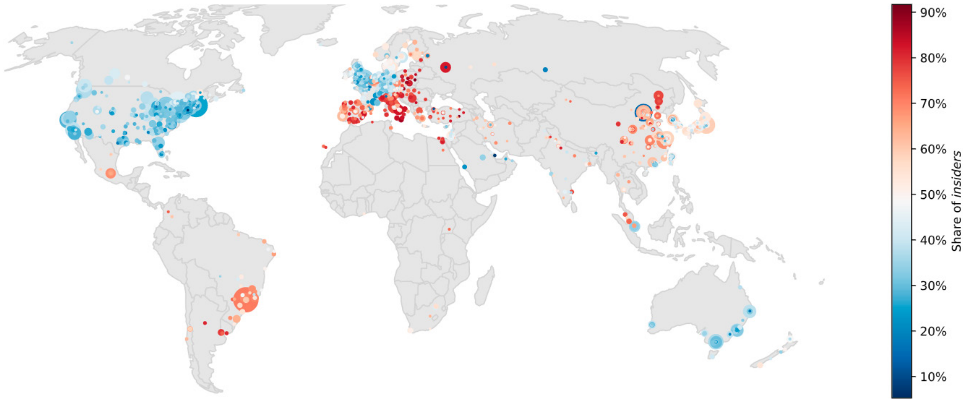


Figure 1. World map of universities included in the dataset and share of insiders per university.
Note: Only universities included in the Leiden Ranking and established before 1998 are included. Nodes denote universities. The size of nodes denotes the number of researchers identified. Colour reflects the share of insiders. Red indicates the highest share while blue reflect lower shares of insiders.

Figure 2 shows the distribution of universities based on the share of insiders in each geographic region. Europe, which is well represented in the sample, is further subdivided into three sub regions (Western and Northern Europe, Southern Europe and Central and Eastern Europe). Traditionally advanced countries in North America, Western and Northern Europe and Oceania attract researchers originating from their own universities. Within North America and Oceania, there is relatively little variation; half of the universities are boxed in a fairly narrow range, indicating that this is a systemic feature of how labour markets for researchers operate therein. Universities in Western and Northern Europe appear to be more diverse, with some high values, particularly in Scandinavian and Benelux countries.

In contrast, a higher share of insiders is typical for universities in former soviet countries from Central and Eastern Europe, as well as in Southern Europe, where the tendency to employ insiders is roughly twice as high as in the Northern and Western parts of the continent. This shows that there are lingering differences within the European Research Area (ERA), even among the "old" member countries. In fact, Southern, Central and Eastern European universities have a propensity to employ insiders that is above of what is common in developing countries, including Africa, although evidence from the latter should be taken with a grain of salt due to a low number of observations from this continent.

For example, only one out of every four researchers at Harvard University, the University of Chicago, the University of Warwick and the Humboldt University of Berlin, but also the University of South Carolina, the Coventry University, or the University of Paderborn, started their career at the same institution. However, more than three out of every four currently affiliated researchers started publishing at Sapienza University of Rome, the University of Seville, the University of Warsaw, the University of Szeged, or the Moscow State University.

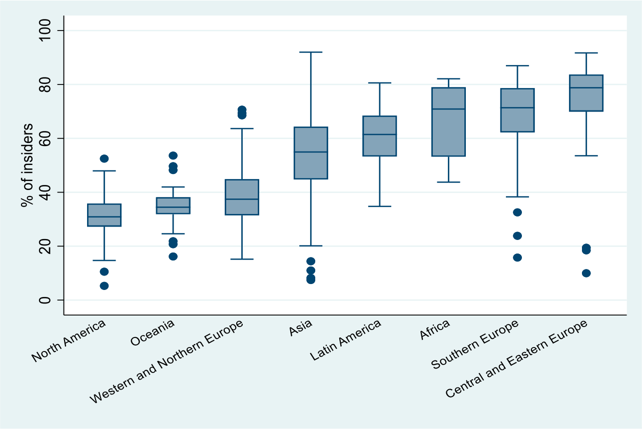


Figure 2. Distribution of the share of insiders in universities grouped by geographic region
Note: Only Leiden Ranking universities that were established before 1998.

When we rank the continents by the median, Latin America and Asia fall between the two extremes. However, this masks vast differences within Asia, which is an amalgam of diverse countries ranging from advanced Japan through emerging China to developing India, but which proves difficult to divide by geography or other lines due to a low number of observations in the potential constituent parts and because the majority of them would be dominated by the largest countries. As illustrated below, the diversity within Asia is best understood by examining evidence from individual countries.

The outliers below the lower limit of the whiskers are predominantly elite local universities that seem not follow the suit in the higher shares of insiders that is common in their national environments. For example, this includes National Research University Higher School of Economics and ITMO University in Russia, Pompeu Fabra University in Spain, University of Cyprus, Qatar University and the University of Chinese Academy of Sciences. On the other end of the spectrum, with the share of insiders above the upper limit of the whiskers, the main outliers include Ghent University and University of Liège in Belgium or Åbo Akademi University in Finland.

Figure 3 complements these patterns by providing details on individual countries, for which results for at least 10 universities are available. Poland comes out clearly at the top of the list: more than four out of five researchers are typically classified as insiders, and none of Polish universities goes below two-thirds of insiders. Next are Italy, Turkey and Spain with below the lower limit of the whiskers lower rates than Poland, but far more variability of the values. The largest variability is detected in China and India; both big countries with emerging and diverse university systems, where the share of insiders ranges from the highest figures in the world to the minimums observed in advanced countries. United States, France, Australia, the United Kingdom and Germany appear at the bottom of the list, which confirms that there is a strong developmental dimension in this ranking.

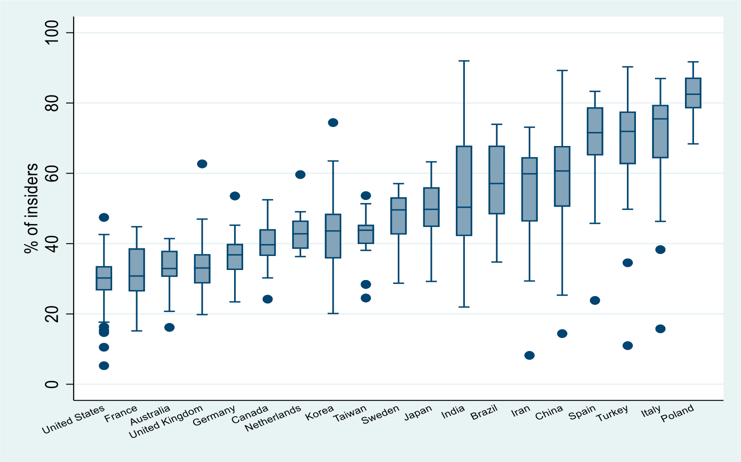


Figure 3. Distribution of insiders in universities grouped by country

Note: Only Leiden Ranking universities that were established before 1998. Only countries with results at least 10 universities.

The share of foreign outsiders is negatively correlated to the size of national research systems. Thus, in large national research systems with many researchers, universities can draw from a large pool of domestic candidates, hence their chances to find a suitable candidate at home are naturally higher than in small research systems with a limited supply. For example, a university in Iceland is just for this reason likely to display a far higher share of foreign outsiders than otherwise similar university operating in the United States. Hence, one should compare in this regard universities from countries with research systems of a roughly similar size.

The largest differences are observed for countries with small research systems. It could be argued that these countries would need to internationalize more their research workforces, particularly if they aim at achieving research excellence in a foreseeable future. However, universities in countries like Slovenia, Croatia, Slovakia, Romania, Lithuania, Algeria, Pakistan and Uganda seem not to attract researchers who started their careers abroad, which is in contrast to Iceland, Cyprus, Lebanon or Jordan. Arabic oil-rich countries in the Persian Gulf, namely Kuwait, Oman, United Arab Emirates, Qatar and Saudi Arabia stand out with high shares of foreign outsiders. This reflects their development strategy based on attracting foreign researchers (Schmoch et al., 2016), but this can also reflect the controversial strategy of universities in Saudi Arabia and elsewhere of offering secondary affiliations to highly cited researchers from abroad to boost their bibliometric profiles and ascend in rankings (Gingras, 2014). Arguably, this reiterates the limitation already mentioned above that we do not have any information on the parameters of contracts that underpin the affiliations of researchers in the published papers, although in practice these researchers contribute to the production (and visibility) of these universities.

Finally, we examine how the results differ by disciplines and thus to which extent these underlying differences could affect the patterns detected above. Using the Dimensions database, we can distinguish between 22 major disciplines across all fields of sciences, including social sciences and humanities. In order to present robust evidence, we narrow the sample only to universities with at least 30 authors in the respective discipline and present the results only for disciplines, for which such data is available in at least 30 universities. On these grounds, we eliminated from the analysis three disciplines (*Law and Legal Studies*; *Philosophy and Religious Studies*; and *Studies in Creative Arts and Writing*). The resulting dataset contains information on 10,408 pairs of university-disciplines of 1,129 universities across 19 disciplines.

Figure 4 displays the boxplots for the share of insiders by disciplines. The main finding is that there is little variability across disciplines but a large diversity within them. The central tendencies are limited to a narrow range, the interquartile ranges are highly overlapping, while the whiskers tend to reach from close to zero to almost hundred percent in most disciplines. Arguably, this picture pales compared to the significant differences that we have observed between universities across the national research systems above.

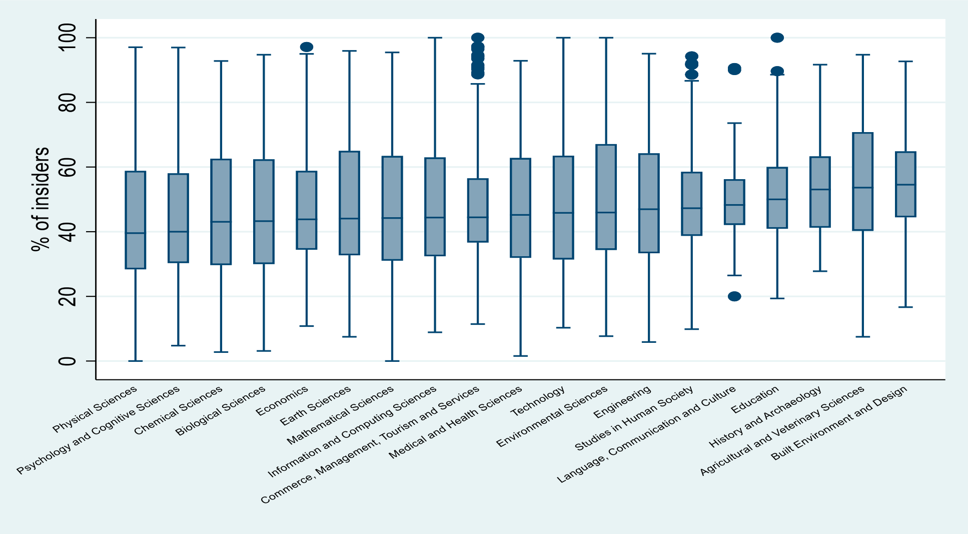


Figure 4. % of insiders by discipline

Note: Only Leiden Ranking universities that were established before 1998 with at least 30 authors in the respective discipline. Only disciplines with results for at least 30 universities.

Conclusions

This paper demonstrates how bibliometric data can provide valuable insights into the institutional mobility of researchers. The empirical analysis reveals noticeable differences along the north versus south and west versus east geographical dimensions. The gulf between universities in Central, Eastern, and Southern Europe on the one hand and universities in North America, Western, and Northern Europe on the other is striking, pointing to systemic differences in how labor markets for researchers operate in the respective areas. Most developing countries fall between the two extremes, but there is significant variation within them. The findings also show that differences between universities and national research systems are most important, while disciplinary differences might be only marginally important.

The main findings on cross-country differences are broadly consistent with the growing body of empirical literature on geographical mobility of researchers, such as the MORE3 survey in Europe conducted by IDEA Consult, WIFO and Technopolis (2017) and the bibliometric analysis of mobility between European countries and the United States (Science Europe, 2013). Nevertheless, evidence on (a lack of) institutional mobility, or more specifically academic inbreeding, has been limited to qualitative analyses, surveys of particular contexts and/or case-studies of individual countries (Inanc and Tuncer 2011, Morichika and Shibayama 2015, Tavares, et al. 2015, Yudkevich et al. 2015, Horta and Yudkevich 2016 and Seeber et al. 2016), whereas broad comparative evidence based on harmonized data has been lacking. In this regard, the approach developed in this paper opens up new avenues for quantitative research as well as policy analyses on this topic.

References

Australian Bureau of Statistics. (2008). Australian and New Zealand Standard Research Classification (ANZSRC). https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1297.02008?OpenDocument

Aref S, Zagheni E, West J. 2019. The Demography of the Peripatetic Researcher: Evidence on Highly Mobile Scholars from the Web of Science In: Weber I, Darwish KM, Wagner C, Zagheni E, Nelson L, Aref S, Flöck F, editors. Social Informatics, Lecture Notes in Computer Science. Cham: Springer International Publishing. pp. 50–65. doi:10.1007/978-3-030-34971-4_4

Berelson, B. (1960). *Graduate Education in the United States*. New York: McGraw-Hill.

Bode, C., Herzog, C., Hook, D., & Mcgrath, R. (2018). A Guide to the Dimensions Data Approach https://doi.org/10.6084/m9.figshare.5783094

Caplow, T., & McGee, R. J. (1958). *The academic marketplace*. Transaction Publishers.

Caron, E., & NJ van Eck. (2014). Large scale author name disambiguation using rule-based scoring and clustering. Available at: . http://www.academia.edu/download/42806714/Research_quality_characteristics_of_publ20160218-10100-1wsawux.pdf#page=91

Cruz-Castro, L., & Sanz-Menéndez, L. (2010). Mobility versus job stability: Assessing tenure and productivity outcomes. *Research policy*, 39(1), 27-38.

European Commision. (2012). Enhancing and focusing EU international cooperation in research and innovation: A strategic approach. http://ec.europa.eu/research/evaluations/index_en.cfm?pg=fp7

European Commission/EACEA/Eurydice. (2015). *The European Higher Education Area in 2015: Bologna Process implementation report*. Luxembourg: Publications Office of the European Union.

Franzoni, C., Scellato, G., & Stephan, P. (2014). The mover's advantage: The superior performance of migrant scientists. *Economics Letters*, *122*(1), 89-93.

Ganguli, I. (2015) Immigration and ideas: What did Russian scientists "bring" to the US? Unpublished manuscript, Stockholm School of Economics

Gringas, Y. (2014). How to boost your university up the rankings. University World News. July 2014 Available at: https://www.universityworldnews.com/post.php?story=20140715142345754.

Gurney, T., Horlings, E., & van den Besselaar, P. (2012). Author disambiguation using multi-aspect similarity indicators. Scientometrics, 91(2), 435–449. https://doi.org/10.1007/s11192-011-0589-1

Hazelkorn E. 2011. Rankings and the Reshaping of Higher Education: The Battle for World-Class Excellence. *Palgrave Macmillan*.

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. Quantitative Science Studies, 1(1), 387-395.

Hoekman, J., 2012. Science in the Age of Globalization, The Geography of Research Collaboration and Its Effect on Scientific Publishing. Eindhoven University of Technology, Eindhoven.

Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: Building Context for Search and Evaluation. Frontiers in Research Metrics and Analytics, 3, 23. https://doi.org/10.3389/frma.2018.00023

Horta, H., Veloso, F. M., & Grediaga, R. (2010). Navel gazing: Academic inbreeding and scientific productivity. *Management Science*, 56(3), 414-429.

Horta, H. (2013) Deepening our understanding of academic inbreeding effects on research information exchange and scientific output: new insights for academic based research. Higher Education, 65, 487-510. http://dx.doi.org/10.1007/s10734-012-9559-7

Horta, H., Yudkevich, M. (2016) The role of academic inbreeding in developing higher education systems: Challenges and possible solutions. Technological Forecasting & Social Change, 113, 363-372. http://dx.doi.org/10.1016/j.techfore.2015.06.039

Inanc, O., Tuncer, O. (2011) The effect of academic inbreeding on scientific effectiveness. Scientometrics, 88, 885–898. http://dx.doi.org/10.1007/s11192-011-0415-9

Jacob, M., & Meek, V. L. (2013). Scientific mobility and international research networks: Trends and policy tools for promoting research excellence and capacity building. Studies in Higher Education, 38(3), 331–344. https://doi.org/10.1080/03075079.2013.773789

IDEA Consult, Wifo, Technopolis (2017). MORE3–Support Data Collection and Analysis Concerning Mobility Patterns and Career Paths of Researchers. *WIFO Studies*.

Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. In Scientometrics (Vol. 90, Issue 2, pp. 463–481). Springer. https://doi.org/10.1007/s11192-011-0495-6

Lee, W. M. (2000). Publication trends of doctoral students in three fields from 1965–1995. Journal of the American Society for Information Science, 51(2), 139–144.

Lepori, B., Seeber, M., & Bonaccorsi, A. (2015). Competition for talent. Country and organizational-level effects in the internationalization of European higher education institutions. *Research policy*, 44(3), 789-802.

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. Journal of the American Society for Information Science and Technology, 63(5), 1030–1047. https://doi.org/10.1002/asi.22621

Macháček, V. and Srholec, M. (2020) Where do universities recruit researchers?, IDEA Study 1/2020. Institute for Democracy and Economic Analysis (IDEA), CERGE-EI, Prague. https://idea.cerge-ei.cz/files/RecruitingResearchers/

Mamiseishvili, K., & Rosser, V. J. (2010). International and citizen faculty in the United States: An examination of their productivity at research universities. *Research in Higher Education*, 51(1), 88.

Melin, G. (2005). The dark side of mobility: negative experiences of doing a postdoc period abroad. *Research Evaluation*, *14*(3), 229-237.

Meyer, J.B. (2001) Network approach versus brain drain: lessons from the diaspora. International migration, 39(5), pp.91-110

Moed, H. F., & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. Scientometrics, 101(3), 1987–2001. https://doi.org/10.1007/s11192-014-1307-6

Morichika, N. Shibayama, S. (2015) Impact of inbreeding on scientific productivity: A case study of a Japanese university department. Research Evaluation, 24, 146-157.

Moguérou, P., & Di Pietrogiacomo, M. P. (2008). Stock, Career and Mobility of Researchers in the EU. JRC Scientific and Technical Reports.

Nane, G. F., Larivière, V., & Costas, R. (2017). Predicting the age of researchers using bibliometric data. Journal of informetrics, 11(3), 713-729.

OECD. (2008). The Global Competition for Talent: Mobility of the Highly Skilled. OECD.

Robinson-Garcia, N., Sugimoto, C. R., Murray, D., Yegros-Yegros, A., Larivière, V., & Costas, R. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. Journal of Informetrics, 13(1), 50–63. https://doi.org/10.1016/j.joi.2018.11.002

Scellato, G., Franzoni, C., & Stephan, P. (2012). Mobile scientists and international networks (No. w18613). *National Bureau of Economic Research*.

Science Europe (2013) Comparative Benchmarking of European and US Research Collaboration and Researcher Mobility. Science Europe and Elsevier's SciVal Analytics.

Schmoch, U., Fardoun, H. M., & Mashat, A. S. (2016). Establishing a World-Class University in Saudi Arabia: intended and unintended effects. Scientometrics, 109(2), 1191-1207.

Seeber, M., Debacker, N., & Vandevelde, K. (2016) Mobility and inbreeding in the heart of Europe. What factors predict academic career in Dutch-speaking Belgian universities. In Paper presented to the OECD blue sky forum on science and innovation indicators

Stephan, P.E., Levin, S.G. (2001) Exceptional contributions to US science by the foreign-born and foreign-educated. Population Research and Policy Review 20, 59–79 https://doi.org/10.1023/A:1010682017950

Sugimoto, C. R., Robinson-Garcia, N., & Costas, R. (2016). Towards a global scientific brain: Indicators of researcher mobility using co-affiliation data. http://arxiv.org/abs/1609.06499

Sugimoto, C. R., Robinson-García, N., Murray, D. S., Yegros-Yegros, A., Costas, R., & Larivière, V. (2017) Scientists have most impact when they're free to move. Nature News, 550(7674), 29

Tavares, O., Cardoso, S., Carvalho, T., Sousa, S.B., Santiago, R. (2015). Academic inbreeding in the Portuguese academia. Higher Education. 69, 991–1006.

Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? Journal of Informetrics, 12(2), 430–435. https://doi.org/10.1016/j.joi.2018.03.006

Van Eck, Nees Jan. (2020). CWTS Leiden Ranking 2020 [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3997240

Waaijer, C. J. F., Macaluso, B., Sugimoto, C. R., & Larivière, V. (2016). Stability and Longevity in the Publication Careers of U.S. Doctorate Recipients. PLOS ONE, 11(4), e0154741. https://doi.org/10.1371/journal.pone.0154741

Wagner, C.S. and Jonkers, K., (2017) Open countries have strong science. Nature News, 550(7674), p.32.

Wagner, C. S., & Leydesdorff, L. (2005). Mapping the network of global science: comparing international co-authorships from 1990 to 2000. *International journal of Technology and Globalisation*, *1*(2), 185-208.

Welch EW, van Holm E, Jung H, Melkers J, Robinson-Garcia N, Mamiseishvili K, Pinheiro D. (2018). The Global Scientific Workforce (GTEC) Framework. STI 2018 Conference. Leiden (The Netherlands): Centre for Science and Technology Studies (CWTS). pp. 868–871.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., van Leeuwen, T. N., van Raan, A. F. J., Visser, M. S., Wouters, P. (2012) The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12), 2419-2432.

Yudkevich, M., & Sivak, E. (2012) University inbreeding: an impact on values, strategies and individual productivity of faculty members. Strategies and Individual Productivity of Faculty Members (January 31, 2012).

Yudkevich, M., Altbach, P. G. and Rumbley, L. E. (2015) Academic Inbreeding and Mobility in Higher Education. Global Perspectives. New York: Palgrave Macmillan.

O Factor: Open-Open citation measure

Devika P. Madalli

devika@drtc.isibang.ac.in
Documentation Research and Training Centre, Indian Statistical Institute, Bangalore (India)

Abstract

Measures such as Impact factor and H-Index are used to reflect the impact of scientific publishing and authors' productivity. There have been many arguments in support of and against these measures: whether they truly taking all factors into consideration and whether they must be used as the bottom-line in deciding worth of published content and authors who contribute content. However, measures such as Impact factor and H-Index and other indices continue to used, rather overwhelmingly, as measures of quality and productivity. The other facet the quality paradigm is the question of quality of Open content. Open access advocates endeavour to entice academicians and researchers to publish in open journals and other open resources. Academicians are hesitant to publish in open access journals because open access publishing has still not attained the status that commercial published journals have. Here we address the question 'How to establish quality of Open Content?' We introduce the concept of 'Open Content Metrics (OCM)' as counterparts of the measures of quality and productivity already established for commercially published journals. We introduce the concept of O–Factor (OF) as a measure to reflect what we refer as Open-Open citation. Data and illustration to calculate OF are provided.

Introduction

Measures of scientific productivity and outputs are important indicators that reflect the progress of science. Citation index is used as a means to attribute to authors their contribution and Impact factor as a means to measure the impact of a journal. These provide a mechanism to assess the impact of a researcher in the scientific community. There are several arguments for and against the appropriateness of such measures (Garfield, 1955; Kelly and Jennions, 2006; Ayris *et al.,* 2018), but measures such as impact factors continue to be used and quoted as indicators for quality of research. Proponents and advocates of open content, work towards making content openly available with appropriate licenses that foster the sharing of knowledge, thus hoping to free it from the commercial publishing cycle. However open content is often discarded with a common rather unfair remark that 'All' Open Content is of poor quality. Academics and researchers who are better judges of quality know that there is good and bad in Open content just as is the case with commercially published content. The main reason being that it is academicians and researchers who produce content but it is also true that the 'same' academics work as peer reviewers to ensure 'quality' of content whether in open access and or in commercially published journals. Other reasons for this are many and are deemed as out of the scope of the present discussion. To sum up, the discussion of open versus commercial journals often culminates in a common unfounded insecurity that open access content lacks in quality. To prove or disprove the argument may be the subject of prolonged discussions but the present work is based on the premise of the Open Access Principles (Madalli, 2011) that state: ***'Knowledge unto those that produce knowledge'*** *and* ***'Scholarship is the same; priced or open'.***

In view of the above, it is important to concentrate on measures that can be applied to Open Content and we refer to such measures as 'OCM – Open Content Metrics'. Open Content metrics are meant to encompass all measures that establish productivity, quality, use and reuse of Open Content. The focus of this paper is to deliberate upon the impact of open access journals and open content and to introduce the concept of 'Open Authors'. 'Open authors' is a term referred to authors publishing in open access journals. The significance of the proposal lies in establishing impact of the Open content and providing the right impetus for researchers and authors to consider being ‘Open Authors' without any concerns or constraints.

Quality and Measures of scientific publishing

There have been several critical studies (Seglen, 1997; Linde, 1998) explicating the pitfalls of the conventional methods of quantifying scientific literature output like the impact factor. Some of the most common pitfalls are enumerated below:

- The overall impact of a journal as quantified by its impact factor is not a marker of the quality of the individual articles of that journal.
- Impact Factor is highly domain-specific, and, thus it is unsuited to be used as a standard to gauge research output across domains.
- It is also plagued by quantitative pitfalls like the number of references attributed or the number of self-citations, which introduces noise in the final calculation of the metric
- Impedances like the language, visibility and the length of the articles within a journal also affects its impact factor, which should ideally not be a concern.
- Due to its dependence on the dynamics of a specific domain, Impact Factor might show undue inflation/deflation in its measured value.
- Also, impact factor cannot be said to be absolutely foolproof as there are byways through which it can be subjected to quantitative manipulation and misrepresented.

Carpenter, Cone and Sarli (2014) elucidate the various upcoming methods of measuring scientific and research output, specifically highlighting the importance of more granular, document-specific metric measures such as counts reflecting events like views, downloads, tags, recommendations and annotations which are taken into consideration in the altmetrics approach. They argue that such measures, as opposed to the conventional methods to gauge scientific outputs, can effectively reflect the inceptive sway and provide a much broader perspective as regards the research importance of a journal and its constituent articles. They also point out that there may be a chance of co-existence amongst the conventional and the next-generational metrics. Barnes (2015) takes a more nuanced approach while considering altmetrics as an innovative new tool in measuring scientific impact. He reviews various pieces of evidence indicating much more timely impact assessment provided by altmetric indicators as well as their wider coverage compared to the traditional metrics. Bornmann (2014) discusses the benefits and pitfalls of alternative metrics of measuring scientific output in an illustrative and engaging way. The major benefits of such alternative metrics are listed to be broadness, diversity, speed and openness.

The above discussion from literature evidence that 1) There are many measures being considered and applied to gauge quality and productivity in scientific publishing. 2) That none are perfect and 3) That devising metrics and methodologies towards measuring productivity and impact of scientific content is an ever emerging domain by itself, with newer modes of publishing and use and reuse of content in novel ways.

Hence while research is required to arrive at the best measures, the academic and scientific world continues to use Impact factor and H, G (and similar) indices as measures to reflect quality. It is necessary to establish measures for open content hosting resources including Open journals to establish themselves as an equally good opportunity to publish. In the sections that follow we introduce the concept of O-factor as a measure of impact of Open access journals.

Data Collection

Collecting data about Open Authors is most challenging in the present situation where commercial citation indices make no distinction about open authors and open content and that of commercial content. Given this constraint, in the first instance, we selected Google Scholar (GS) as our citation database to collect citation data. To identify information whether a given author has published in open access we proceed to the next step. GS provides citation information and other bibliographic information. For finding the citation information of an article cited by Open Access journals we used the citation information link in GS where we checked which sources have cited the article and collected only those citations given by open access journals. However, there were a few problems to be addressed:

- GS indexes different types of resources including scholarly journals as well as web-based resources. So, when we searched for an article, we found the citation count was more compared to Scopus because it collects citation information from all the resources like websites, slide-share, etc.

- Next problem was, GS collects citation from different languages including English, so when we analyze the citation information, it is difficult for us to identify whether the cited document is from open access journal or from other resources.

To overcome these problems, we have used Scopus for our study. To validate our concept we have considered an Open Access (OA) journal, 'Journal of Medical Library Association (JMLA)' from Directory of Open Access Journal (DOAJ). This title is a random selection and with an aim to validate our idea and illustrate O factor as one measure within OCMs to establish the impact of Open Content.

To collect citation-related information, we have used the search keyword "Journal of Medical Library Association" and restricted the datasets to the year 2016 and 2017 because we are trying to find the impact factor and O-factor for the year 2018. Based on the above query we found 142 articles from Scopus for the years 2016 and 2017.

Then, we searched each article in the Scopus database and collected citation count as well as the citation given by the OA journals. As is obvious, this is done by manual verification. For our study we only considered five fields viz. article title, year of publication, source of the article, cited by and cited by OA journals.

One of the problems we faced during collecting the citation information from Scopus was, some of the articles which we collected did not provide complete citation information i.e., it had citation count but did not provide information about who cited those articles.

***Table-1: Excerpt of Citation* Information of Journal of Medical Library Association (JMLA) for 2017 and 2016**

Sr.No.	*Article Title*	*Year*	*Source title*	*Cited by*	*Cited by OAJ*
1	Facing reality: The growth of virtual reality and health sciences libraries	2017	JMLA	0	0
2	Examining the role of MEDLINE as a patient care information resource: An analysis of data from the value of libraries study	2017	JMLA	2	1
3	PsycEXTRA	2017	JMLA	0	0
4	Enhancing the research and publication efforts of health sciences librarians via an academic writing retreat	2017	JMLA	2	2
5	Response to "phrase truncation in PubMed searches"	2017	JMLA	0	0
6	Characteristics of multi-institutional health sciences education research: A systematic review	2017	JMLA	0	0
7	Awareness, adoption, and application of the association of college & research libraries (ACRL) framework for information literacy in health sciences libraries	2017	JMLA	4	0
8	2016 audited schedule of changes in net assets	2017	JMLA	0	0
9	Barbara Epstein, AHIP, FMLA, medical library association president, 2017-2018	2017	JMLA	0	0
10	Teaching evidence-based practice principles to prepare health professions students for an interprofessional learning experience	2017	JMLA	2	1
.	.....	....	....	......	......
.	....	....	......	......	......
.	......	.....	....	....	....
141	Rural providers' access to online resources: A randomized controlled trial	2016	JMLA	4	1
142	Evolution of biomedical communication as reflected by the National Library of Medicine	2016	JMLA	0	0
	Total			**311**	**107**

Data Analysis

Based on the data collected from Scopus in table 1, we have calculated IF and O-Factor with the help of three year IF formula which was given by Garfield (Garfield, 2006), namely:

Computing the Impact Factor (IF)

Impact factor of a journal is calculated taking into consideration citations received recently which is defined as 'in the two years that precede the year of calculation of its IF'. This is devised to assess the Impact as well as to maintain the current status that helps the current ranking of journals.

IF = Total number of citations received by journal for the previous two years/
Total number of articles published in previous two years

Computing the O factor (OF)

We introduce the notion of 'O' factor (OF), which one of OCMs. OF is derived by computing citations received by an open access journal. O factor for an open access journal is defined as:

'the total number of the citations received by articles in the journal by open authors divided by the total number of articles published in the open access journal in the preceding two years of its publication'

In a given year 'y' OF of an OA journal is computed as

$$OF_y = C_{y-1} + C_{y-2} / P_{y-1} + P_{y-2}$$

where
OF is the Open Impact Factor
y = the year for which impact factor is calculated
C = citations received
P = total no. of publications in the journal in a given year (y-1, y-2 ... so on)

Illustration

In our illustration, for the JMLA we found that the total number of citations received in the year 2016 and 2017 was 311 (see Appendix I for full data).
The total number of articles published in the year 2016 and 2017 was 142.

So, the Impact Factor of JMLA for the year 2018 is

$$IF= 311/142=2.19$$

For finding O-Factor (OF) of JMLA of 2018,

$$OF_{2018} = C_{2017} + C_{2016} / P_{2017} + P_{2016}$$

where
OF is the Impact Factor
y = the year for which impact factor is to be calculated = 2018
C = citations received from open access journal only = 107
P = total no. of publications in the journal in the year 2016 and 2017 = 142
so,

$$\text{O-Factor}_{2018} = 107/ 142 = 0.75$$

Findings

From the above analysis, we found that IF of JMLA is 2.19 and the O-Factor of JMLA is 0.75. The result shows that IF is more than the O-factor but if we consider the total number of citations received from OA journals, it was around 36% of the total citation received from all the journals. The other finding is, the concept of O-Factor is new and there are many constraints to find the O-Factor of a journal at present. For instance, if a journal publishes 1000 articles in two years then getting all the required information to calculate O-Factor is a very difficult task. It is hoped that OCMs and methodologies for deriving OCMs will be refined and will be used as commonly as IF, H index and other measures in future.

Limitations

The concept of OCM is nascent and literature reveals very scant information on efforts to work out measures for open content. Popular citation indices do not distinguish between commercial content and open content and there are no indicators for articles published in Open access journals or other open resources. It takes manual effort to identify first which open access journals and authors have gained citations, and second whether the citation is referring to openly published content or commercial content or a mix of both. For the purpose of illustrating OF, we have used data from Scopus but we do not make a presumption to what extent Scopus covers open access content. We have just used one journal, 'JMLA' and collected data for the specified 2 years to demonstrate the methodology. Data as well as findings may vary rather drastically by the domain, title of publication, country of publication and other such factors.

Significance and Conclusion

As discussed, and illustrated above, there are ways to establish 'Quality' by existing and accepted norms of measuring the impact of scientific publishing. The existing and accepted measures have their pitfalls as evidenced within this paper but to establish the worth of open content and encourage authors to consider being "Open Authors" it is necessary to establish measures that work as indicators for their productivity. Such measures will also enable open access journals to be listed and considered administratively for the purposes of career advancement in academia on par with so-called "Ranked journals", which of course has been the crux of the matter of the reason why authors are hesitant to publish in Open access journals. The notion of 'Open Content Metrics - OCM' is nascent and needs the attention of academia, researchers and open access advocates to bring out effective methodologies and measures in favour of open content. There is no doubt that qualitative indicators such as usage patterns of open access journals and journals articles as such have their merit. The latest metrics such as altmetrics that take into consideration social media resources along with others are also important in reflecting use and reuse of resources. OCMs ideally would encompass all such measures that establish 'quality' of open content.

References

Ayris, P. *et al.* (2018). *Open Science and its role in universities: A roadmap for cultural change* (Vol. 24, pp. 18-19, Advice paper). LERU.

Barnes, C. (2015). The use of altmetrics as a tool for measuring research impact. *Australian Academic & Research Libraries,* 46(2), 121-134, DOI: 10.1080/00048623.2014.1003174

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics,* 8(4), 895-903. doi:10.1016/j.joi.2014.09.005

Carpenter, C. R., Cone, D. C. & Sarli, C. C. (2014). Using Publication Metrics to Highlight Academic Productivity and Research Impact. *Academic Emergency Medicine,* 21(10), 1160-1172. doi:10.1111/acem.12482

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science,* 122(3159), 108-111. doi:10.1126/science.122.3159.108

Garfield, E. (2006). The history and meaning of the Journal Impact Factor. *JAMA*, 295(1), 90-93. https://doi.org/10.1001/jama.295.1.90

Kelly, C. & Jennions, M. (2006). The h index and career assessment by numbers. *Trends in Ecology & Evolution,* 21(4), 167-170. doi:10.1016/j.tree.2006.01.005

Linde, A. (1998). On the pitfalls of journal ranking by Impact Factor R. *European Journal of Oral Sciences,* 106(1), 525-526. doi:10.1046/j.0909-8836..t01-1-.x

Madalli, D.P. (2011). Report on Asia and Pacific for Global Open Access Portal, UNESCO. (unpublished).

Seglen, P. O. (1997). Citations and journal impact factors: Questionable indicators of research quality. *Allergy,* 52(11), 1050-1056. doi:10.1111/j.1398-9995.1997.tb00175.x

Appendix

Table-1 *Citation* Information of Journal of Medical Library Association (JMLA) for 2017 and 2016.

Sr.No.	Title	Year	Source title	Cited by	Cited by OAJ
1	Facing reality: The growth of virtual reality and health sciences libraries	2017	JMLA	0	0
2	Examining the role of MEDLINE as a patient care information resource: An analysis of data from the value of libraries study	2017	JMLA	2	1
3	PsycEXTRA	2017	JMLA	0	0
4	Enhancing the research and publication efforts of health sciences librarians via an academic writing retreat	2017	JMLA	2	2
5	Response to "phrase truncation in PubMed searches"	2017	JMLA	0	0
6	Characteristics of multi-institutional health sciences education research: A systematic review	2017	JMLA	0	0
7	Awareness, adoption, and application of the association of college & research libraries (ACRL) framework for information literacy in health sciences libraries	2017	JMLA	4	0
8	2016 audited schedule of changes in net assets	2017	JMLA	0	0
9	Barbara Epstein, AHIP, FMLA, medical library association president, 2017-2018	2017	JMLA	0	0
10	Teaching evidence-based practice principles to prepare health professions students for an interprofessional learning experience	2017	JMLA	2	1
11	The role of librarians in teaching evidence-based medicine to pediatric residents: A survey of pediatric residency program directors	2017	JMLA	0	0
12	Epistemonikos	2017	JMLA	0	0
13	BrowZine	2017	JMLA	0	0
14	Between the sheets	2017	JMLA	0	0
15	Phrase truncation in PubMed searches	2017	JMLA	0	0
16	Characteristics of personal health information management groups: Findings from an online survey using Amazon's mTurk	2017	JMLA	0	0
17	Advancing the conversation: Next steps for lesbian, gay, bisexual, trans, and queer (LGBTQ) health sciences librarianship	2017	JMLA	2	0
18	Collaboration challenges in systematic reviews: A survey of health sciences librarians	2017	JMLA	2	1
19	Rapid transformation of two libraries using Kotter's Eight Steps of Change	2017	JMLA	0	0
20	Locating sex- and gender-specific data in health promotion research: Evaluating the sensitivity and precision of published filters	2017	JMLA	0	0
21	Cultivating a community of practice: The evolution of a health information specialists program for public librarians	2017	JMLA	2	0
22	Our journey to digital curation of the Jeghers Medical Index	2017	JMLA	0	0
23	Updating search strategies for systematic reviews using endnote	2017	JMLA	1	0
24	History matters…through partnerships that advance research, education, and public service	2017	JMLA	0	0
25	Surveying hospital nurses to discover educational needs and preferences	2017	JMLA	1	0
26	Trend analysis of journal metrics: A new academic library service?	2017	JMLA	0	0

27	A competency framework for librarians involved in systematic reviews	2017	JMLA	1	0
28	VetCompanion: Response and clarifications	2017	JMLA	0	0
29	Native Voices: Native Peoples' Concepts of Health and Illness in New Mexico: Opening a local conversation by hosting a national traveling exhibit	2017	JMLA	0	0
30	Introducing altmetrics to the JMLA	2017	JMLA	0	0
31	Ethics in academic publishing: A timely reminder	2017	JMLA	3	2
32	Using scenario-based training to promote information literacy among on-call consultant pediatricians	2017	JMLA	1	0
33	Compliance of systematic reviews in veterinary journals with Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) literature search reporting guidelines	2017	JMLA	4	0
34	Data Day to Day: Building a community of expertise to address data skills gaps in an academic medical center	2017	JMLA	6	5
35	Creating a web-based digital photographic archive: One hospital library's experience	2017	JMLA	1	1
36	Evaluation of hospital staff's perceived quality of librarian-mediated literature searching services	2017	JMLA	0	0
37	Anticipating the third century of the National Library of Medicine: A call for engagement	2017	JMLA	0	0
38	Correction to: 116th Annual Meeting, Medical Library Association, Inc., Toronto, ON, Canada, May 15–20, 2015 (J Med Libr Assoc, (2017), 105, 1 (E1–E20), 10.5195/jmla.2017.123)	2017	JMLA	0	0
39	Body of evidence: Integrating Eduard Pernkopf's Atlas into a librarian-led medical humanities seminar	2017	JMLA	1	1
40	Investigating the need for scholarly communications positions in Association of Academic Health Sciences Libraries member institutions	2017	JMLA	1	1
41	Interinstitutional collaboration for end-user bioinformatics training: Cytoscape as a case study	2017	JMLA	2	2
42	Mapping the Association of College and Research Libraries information literacy framework and nursing professional standards onto an assessment rubric	2017	JMLA	2	0
43	Whither sir William?	2017	JMLA	0	0
44	Improving data collection, documentation, and workflow in a dementia screening study	2017	JMLA	3	3
45	Publishing case studies in health sciences librarianship	2017	JMLA	2	1
46	Knowledge of journal impact factors among nursing faculty: A cross-sectional study	2017	JMLA	0	0
47	Another one bites the dust…	2017	JMLA	0	0
48	Culturally competent library services and related factors among health sciences librarians: An exploratory study	2017	JMLA	3	2
49	Writing Together to Get AHEAD: An interprofessional boot camp to support scholarly writing in the health professions	2017	JMLA	4	1
50	Evidence-based information needs of public health workers: A systematized review	2017	JMLA	6	2
51	The research life cycle and the health sciences librarian: Responding to change in scholarly communication	2017	JMLA	1	0
52	Correction to De-duplication of database search results for systematic reviews in EndNote (J Med Libr Assoc, (2016) 104, 3, (240-243), 10.3163/1536-5050.104.3.014)	2017	JMLA	0	0
53	Questioning reliability assessments of health information on social media	2017	JMLA	3	2
54	Reviewing retrieved references for inclusion in systematic reviews using endnote	2017	JMLA	7	3

55	Meeting at the crossroads: Collaboration between information technology departments and health sciences libraries	2017	JMLA	1	0
56	Being critical and constructive: A guide to peer reviewing for librarians	2017	JMLA	0	0
57	A day in the life of third-year medical students: Using an ethnographic method to understand information seeking and use	2017	JMLA	2	1
58	116th Annual Meeting, Medical Library Association, Inc., Toronto, ON, Canada, May 15-20, 2015	2017	JMLA	0	0
59	Collection-based analysis of selected medical libraries in the Philippines using doody's core titles	2017	JMLA	0	0
60	Implementing a 3D printing service in a biomedical library	2017	JMLA	2	1
61	Unanswered clinical questions: A survey of specialists and primary care providers	2017	JMLA	6	2
62	We can be heroes: MLA's leadership journey(s)	2017	JMLA	0	0
63	History matters	2017	JMLA	1	0
64	Correction to Creating a library holding group: an approach to large system integration	2017	JMLA	0	0
65	Your teaching strategy matters: How engagement impacts application in health information literacy instruction	2017	JMLA	2	0
66	From English to Chinese, Japanese, and Russian: Extending research visibility with language translations of a conference slide presentation	2017	JMLA	0	0
67	Measuring the health literacy of the upper midwest	2017	JMLA	2	0
68	Erratum to The medical library association guide to answering questions about the affordable care act (J Med Libr Assoc., (2016) 104(3), (248-294), 10.3163/1536-5050.104.3)	2016	JMLA	0	0
69	Evaluation of a health sciences internship for Latino and Native American library students	2016	JMLA	0	0
70	A journal cancellation survey and resulting impact on interlibrary loan	2016	JMLA	1	0
71	Increasing number of databases searched in systematic reviews and meta-analyses between 1994 and 2014	2016	JMLA	6	5
72	Hinari access to research in health program networks to sustain and expand success	2016	JMLA	0	0
73	Manual search approaches used by systematic reviewers in dermatology	2016	JMLA	2	1
74	Impact of librarians on reporting of the literature searching component of pediatric systematic reviews	2016	JMLA	8	6
75	Supplementary searches of PubMed to improve currency of MEDLINE and MEDLINE in-process searches via Ovid	2016	JMLA	5	3
76	How to identify existing literature on patients' knowledge, views, and values: The development of a validated search filter	2016	JMLA	3	2
77	2015 AUDITED SCHEDULE OF CHANGES IN NET ASSETS	2016	JMLA	0	0
78	Sleeping beauties in pediatrics	2016	JMLA	2	1
79	Scoping reviews: Establishing the role of the librarian	2016	JMLA	8	4
80	Research data services in veterinary medicine libraries	2016	JMLA	1	1
81	Using the journal BMJ case reports to promote the publication of clinical case reports	2016	JMLA	0	0
82	Editor's response	2016	JMLA	0	0
83	Mapping the literature of pediatric nursing: Update and implications for library services	2016	JMLA	1	0
84	Tooling up to facilitate findability, virtual collaboration, and storytelling with data	2016	JMLA	0	0

85	Evaluation of resources for analyzing drug interactions	2016	JMLA	8	1
86	Evidence ladder and the JMLA	2016	JMLA	0	0
87	Creating content marketing for libraries	2016	JMLA	0	0
88	Creating a library holding group: An approach to large system integration	2016	JMLA	0	0
89	Equal authorship for equal authors: Personal experience as an equal author in twenty peer-reviewed medical publications during the last three years	2016	JMLA	1	1
90	Let’s get a stronger evidence base for our decisions	2016	JMLA	1	1
91	Integrating research into practice	2016	JMLA	1	0
92	Patron perception and utilization of an embedded librarian program	2016	JMLA	4	1
93	Take us to the beach	2016	JMLA	0	0
94	Virtual embedded librarianship program: A personal view*	2016	JMLA	2	0
95	Resource format preferences across the medical curriculum	2016	JMLA	2	1
96	An elemental strategy	2016	JMLA	0	0
97	Norming a value rubric to assess graduate information literacy skills	2016	JMLA	2	1
98	Effectiveness of adverse effects search filters: Drugs versus medical devices	2016	JMLA	2	2
99	Teresa L. Knott, AHIP, Medical Library Association President, 2016-2017	2016	JMLA	0	0
100	De-duplication of database search results for systematic reviews in endnote	2016	JMLA	23	11
101	What’s new in abstracts of science articles?	2016	JMLA	4	0
102	Sarah Cole Brown, AHIP, FMLA, 1911–2015	2016	JMLA	0	0
103	Librarians help high school students improve research skills	2016	JMLA	0	0
104	Health sciences librarians off the radar	2016	JMLA	1	0
105	Corrections: Epstein BA. In their own words: Oral histories of Medical Library Association past presidents.(J Med Libr Assoc., (2016)Jan, 104(1), (3-14), 10.3163/1536-5050.104.1.002)	2016	JMLA	0	0
106	Instructional methods used by health sciences librarians to teach evidence-based practice (EBP): A systematic review	2016	JMLA	12	0
107	How do early career health sciences information professionals gain competencies?	2016	JMLA	5	0
108	Mapping the literature of hospital pharmacy	2016	JMLA	2	0
109	How to use survey results	2016	JMLA	0	0
110	Finding alternatives when a major database is gone	2016	JMLA	2	1
111	Research survey suggests opportunities	2016	JMLA	1	1
112	Flipping one-shot library instruction: Using Canvas and Pecha Kucha for peer teaching	2016	JMLA	13	6
113	Retraction policies: Standardization?	2016	JMLA	1	0
114	Gaps in affiliation indexing in Scopus and PubMed	2016	JMLA	3	0
115	Comparison of three web-scale discovery services for health sciences research	2016	JMLA	9	0
116	Placing wireless tablets in clinical settings for patient education	2016	JMLA	4	0
117	Research protocols	2016	JMLA	0	0

118	Variation in number of hits for complex searches in Google Scholar	2016	JMLA	4	2
119	Access to human, animal, and environmental journals is still limited for the One Health community	2016	JMLA	1	1
120	Provider documentation of patient education: A lean investigation	2016	JMLA	0	0
121	Research engagement of health sciences librarians: A survey of research-related activities and attitudes	2016	JMLA	10	3
122	Examining care navigation: Librarian participation in a team-based approach?	2016	JMLA	2	0
123	Fast five survey results and response	2016	JMLA	0	0
124	Research submission categories	2016	JMLA	1	1
125	New journals for publishing medical case reports	2016	JMLA	16	6
126	Major bibliographic errors in PubMed: Personal experience among 240 publications and proposed remediation process for errors	2016	JMLA	1	0
127	Scanning technology selection impacts acceptability and usefulness of image-rich content	2016	JMLA	0	0
128	Evaluating the appropriateness of electronic information resources for learning	2016	JMLA	1	1
129	Limits of search filter development	2016	JMLA	4	1
130	Simple export of journal citation data to Excel using any reference manager	2016	JMLA	0	0
131	In their own words: Oral histories of Medical Library Association past presidents	2016	JMLA	0	0
132	National library of medicine response	2016	JMLA	1	0
133	An anniversary and you can help	2016	JMLA	0	0
134	Evaluation of popular drug information resources on clinically useful and actionable pharmacogenomic information	2016	JMLA	3	1
135	Proceedings, 115th Annual Meeting, Medical Library Association, Inc. Austin, TX	2016	JMLA	0	0
136	What is a “mapping study?”	2016	JMLA	14	0
137	Evolution of an academic-public library partnership	2016	JMLA	7	1
138	Mapping studies	2016	JMLA	4	0
139	Performance of a mixed filter to identify relevant studies for mixed studies reviews	2016	JMLA	3	0
140	Data literacy training needs of biomedical researchers	2016	JMLA	10	4
141	Rural providers’ access to online resources: A randomized controlled trial	2016	JMLA	4	1
142	Evolution of biomedical communication as reflected by the National Library of Medicine	2016	JMLA	0	0
Total				311	107

Article Processing Charges based publications: to which extent the price explains scientific impact?

Abdelghani Maddi[1] and David Sapinho[2]

[1] *abdelghani.maddi@hceres.fr*
Observatoire des Sciences et Techniques, Hcéres, 2 Rue Albert Einstein, Paris, 75013 France.
Sorbonne Paris Nord, CEPN, UMR-CNRS 723, Villetaneuse, 93420 France.

[2] *david.sapinho@hceres.fr*
Observatoire des Sciences et Techniques, Hcéres, 2 Rue Albert Einstein, Paris, 75013 France

Abstract

The present study aims to analyze relationship between Citations Normalized Score (NCS) of scientific publications and Article Processing Charges (APCs) amounts of Gold Open access publications. To do so, we use APCs information provided by OpenAPC database and citations scores of publications in the Web of Science database (WoS). Database covers the period from 2006 to 2019 with 83,752 articles published in 4751 journals belonging to 267 distinct publishers. Results show that contrary to this belief, paying dearly does not necessarily increase the impact of publications. First, large publishers with high impact are not the most expensive. Second, publishers with the highest APCs are not necessarily the best in terms of impact. Correlation between APCs and impact is moderate. Otherwise, in the econometric analysis we have shown that publication quality is strongly determined by journal quality in which it is published. International collaboration also plays an important role in citations score.

Introduction

Since the start of the 21st century, scientific community has witnessed an unprecedented rise in the Open Access (OA) movement (Björk, 2004; Chi Chang, 2006; Sotudeh and Estakhr, 2018). OA is seen as a good means of ensuring better dissemination of knowledge and more equity between actors, faced with the issue of paying subscription fees (Prosser, 2003; Tananbaum, 2003; Solomon and Björk, 2012; Cary and Rockwell, 2020).

However, OA does not necessarily mean "free", and raises the question of business model underlying scientific publication. For institutions, it may even generate new costs: in addition to the subscription fees, they are increasingly led to pay costs of OA publication (Maddi, 2020). This concerns a part of Gold OA publications which are based on the "author-pays" business model (Rizor and Holley, 2014; Sotudeh, Ghasempour and Yaghtin, 2015). Thus, authors pay the "Article Processing Charges" (APCs), usually via their institution, to allow open access to the publication (Asai, 2019; Khoo, 2019; Bruns, Rimmert and Taubert, 2020; Copiello, 2020). The APCs have increased significantly over time. This rise has been estimated at three times faster than it would be if indexed to inflation (Khoo 2019). The trend appears to be stronger in more frequently cited journals, as highlighted for Biomed Central journals (Asai 2020), medical and specific OA journals (Asai 2019). These findings would also suggest that large subscription journal publishers tend to set higher APCs. Nevertheless, there is no evidence to date that the introduction of APCs for a given journal reduced its publications volume (Khoo, 2019). In other words, once able to pay an APC, authors give little emphasis to their amount. The APC-based publishing model is being more and more integrated by academic institutions that aspire to switch to an economic model excluding subscription fees. Thus, APCs are now a considerable burden on "total cost of publication" for institutions, reaching 10% in 2013 (Pinfield, Salter and Bath, 2016).

[1] Corresponding author : Abdelghani Maddi, Observatoire des Sciences et Techniques, Hcéres, 2 Rue Albert Einstein, Paris, 75013, France, T. 33 (0)1 55 55 61 48.

While many studies have described relationships between OA publications and citation level (Gumpenberger, Ovalle-Perandones and Gorraiz, 2013; Zhang and Watson, 2017; Piwowar *et al.*, 2018), only few have focused on the amount of APCs, with heterogeneous findings. On one hand, APCs based journals have been regarded, in general, more cited than other OA journals (Björk and Solomon, 2012) , on the other hand, it was concluded that both categories have, on average, similar performances with some disciplinary differences (Ghane, Niazmand and Sabet Sarvestani, 2020). Another study analyzed the relationship between APCs and scientific impact of publications using respectively DOAJ and Scopus data (Björk and Solomon, 2015). On a set of 61,081 publications and 595 journals, authors showed that there is a moderate correlation (0.4) between the two indicators at the journal level (APCs and impact). Correlation is greater (0.6) when data is weighted by the volume of articles for each journal (article level), suggesting that publishers take quality into account when pricing their journals. Likewise, authors are also sensitive to journal quality in their submission choices.
The present study aims to analyze relationship between citations normalized score of scientific publications and APCs amounts. To do so, we use APCs information provided by OpenAPC database (https://treemaps.intact-project.org/) and citations scores in the Web of Science database (WoS). Database covers the period from 2006 to 2019 with 109,141 publications (March 1, 2020). Among these publications, 83,752 match with the WoS database using DOI. Our database contains 4,751 journals with a contrasted number of publications per journal. These journals belong to 267 distinct publishers.
To the extent that large, high impact publishers/journals may request high APCs, it is expected that quality will be strongly correlated with the APCs. The latter would therefore explain the publications visibility.

Data

APCs data

APCs data has been extracted from “OpenAPC” database. It is an initiative that involves 231 institutions worldwide (5 from North America, 255 from Europe and 1 from East Asia) that publish data sets on fees paid for OA journal articles under an open database license. At the beginning of March 2020, the database contains 109,141 publications and 6,941 journals.
Each publication in this database is only assigned to the institution that declared, skipping, therefore, other collaborating institutions. For more details about Open APC database see: https://treemaps.intact-project.org/page/about.html

OST data

The data about citations scores and disciplinary assignation of publications has been extracted from the french Observatoire des Sciences et Techniques' (OST) in-house database. It includes five indexes of WoS available from Clarivate Analytics (SCIE, SSCI, AHCI, CPCI-SSH and CPCI-S) and corresponds to WoS content indexed through the end of March 2019. See https://clarivate.com/webofsciencegroup/solutions/webofscience-platform/.

Final database

The database used for analysis includes several information about publications from the two data sources:

- From the APC data : institution that declares the publication, APCs amounts, journal in which they are published, country of the journal, publisher of journal and a flag indicating whether journal is hybrid. By matching the "OpenAPC" database to that of OST.

- From the WoS-OST in-house data : we estimate the impact of publications by calculating the following indicators :
    - Normalized Citations Score (NCS) at article level: the NCS of a given article was calculated by dividing the number of citations received by the average number of citations in the same disciplines and the same year (Waltman *et al.*, 2011).
    - Mean Normalized Citation Score (MNCS) at journal and publisher level: the MNCS for a given publisher was calculated as the weighted average of the NCS scores, based on all the articles of journals that it publishes. In the case of the OpenAPC database, the selection was restricted to OA articles only.
    - Mean Normalized Impact of Journals (MNIJ) at journal and publisher level: For a given journal, the MNIJ is calculated as the average number of citations per article in a given discipline for a given year, normalized by the number of citations per article in the same discipline and the same year at the global level. The overall MNIJ of a journal (all disciplines combined) is obtained by calculating a weighted average of the MNIJs by discipline.
- Finally, international collaboration was measured by number of countries involved in publication.

The two datasets were merged on the basis of DOI, resulting in a sample of 83,752publications.

Method

Spearman correlation test and regression analysis are performed to highlight the relationships between amount of APC and citations.

Dependent variable and model choice

The dependent variable is the logarithm of Normalized Citations Score (labelled Log (NCS)) received by each publication during the period 2006-2019. To retain the zeros, we have added 1 to the NCS before making the logarithmic transformation. Log (NCS) is a continuous variable with a lower boundary at zero and an upper boundary at infinity. Thus, a left censored Tobit regression model is used to account for the disproportionate number of observations with zero values, because a significant proportion of the observations in our sample are zeros. Tobit regressions avoid inconsistent estimates from OLS regression.

Explanatory variable

In this study, we seek to analyze to what extent the amount of APCs have an incidence on the number of citations received by OA scientific publications. Our explanatory variable is therefore the amount of APCs by publication.

Control variables

Journal impact and number of countries per publication are added as control variables. This choice was driven by the literature that shows that citations depend on journal quality and international collaboration (Maddi, Larivière and Gingras, 2019; Maddi and Gingras, 2021). The hybrid status of a journal was also observed, using a dummy variable.

Results

In this section we present the main results. First, we characterize the APCs data, namely: evolution of the average amount paid by institutions, characteristics of the top 20 producing publishers and then those of the most expensive one. Secondly, we present the correlation tests results between the amount of APCs on the one hand and the MNCS and MNIJ indicators on the other hand. Finally, we present results from regression.

Overview of APCs data

Figure 1 shows the evolution of APCs average by publication in all "OpenAPC" database, and using a constant data set of publications from journals of 2006-09 period (79 journals). As we can see, the average amount has doubled between 2006 and 2019, going from 1,000 euros to almost 2,300 euros per publication. This is explained in particular by the new journals that have been indexed into OpenAPC database after 2009, which significantly increases the APCs average. Overall, APCs have increased significantly even for old journals (2006-09) from 1,000 to 1,800 euros on average.

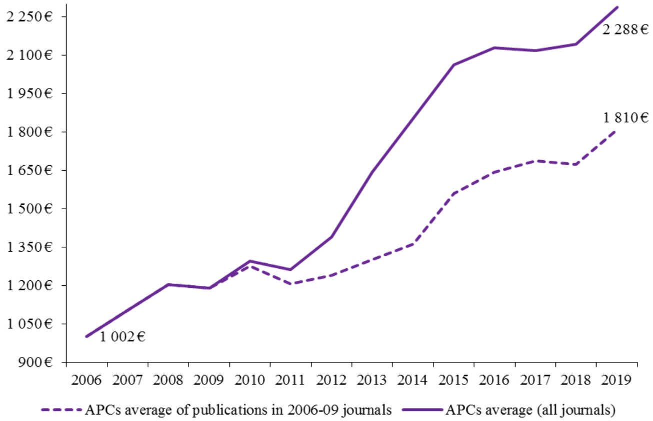


Figure 1: APCs Average per publication, total and with constant journal set

With all reservations that we can make on the Open APC database, we can hypothesize that this increase is notably due to a rise in demand for OA publication. As is well known in economics, increase in demand systematically leads to an increase in prices. Publishers are therefore taking advantage of this enthusiasm for OA to increase their prices.

Figure 2 describes the distribution of publications in the OpenAPC database, in relation to the amount of APCs, by discipline.

First, it highlights contrasting gold open access publication practices. This finding is similar to that observed from the WoS data base for OA publications in previous studies (Maddi, 2020; Maddi, Lardreau and Sapinho, 2021). However, the distribution is quite different than that of all the publications where the weight of engineering or chemistry, for example, is higher (OST, 2019). Thus, nearly 50% of the publications in the OpenAPC database are in Fundamental biology and Medical research, with respectively 23,466 and 19,669 publications. These two disciplines account for only 30% of publications in the WoS database (and 42% of Gold open access). The least present disciplines are mathematics, computer science and humanities. This result is largely explained by the overall size of these disciplines (see OST, 2019).

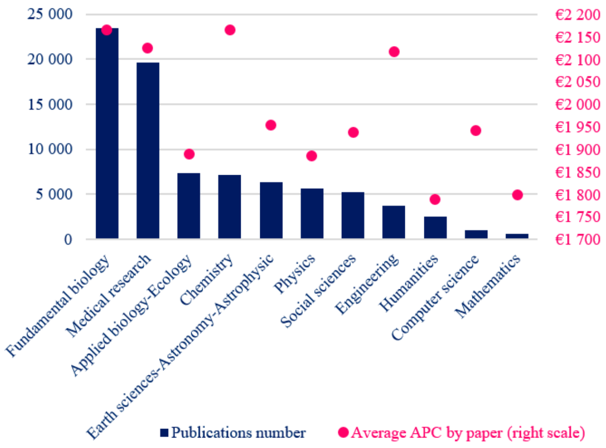


Figure 2: publications number and average APC by discipline

The average amount of APCs by discipline varies from € 1,800 (Mathematics and Humanities) to € 2,150 (Fundamental biology and Chemistry). The difference is therefore not significantly high depending on the discipline.

Figure 3 shows the distribution of publications in the OpenAPC database (top 20 countries) as well as the corresponding average APC amount per country.

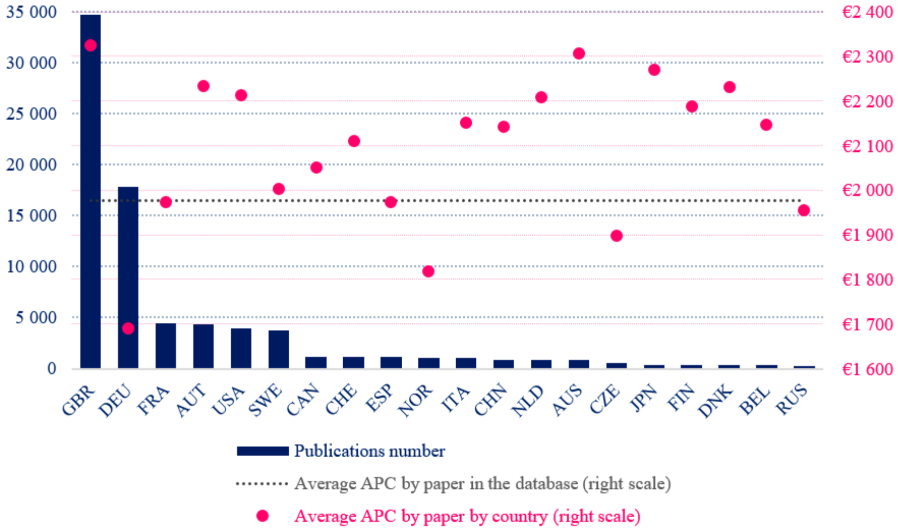


Figure 3: publications number and average APC by country (Top 20 countries in OpenAPC database)

The UK accounts for 40% of the OpenAPC database publications with almost 35,000 articles (fractional count), followed by far by Germany with 17,779 articles. France comes in third position, with 4,450 publications, followed closely by Austria, the United States and Sweden. This distribution shows that the OpenAPC database is biased in favor of European countries. The scope of our results is therefore limited in particular to European countries. When it comes to the average amount of APC per country, the UK also comes first with an average APC of

2325 euros. In contrast, Germany, the second producer country in the OpenAPC database, has the lowest average APC per paper (1690 euros), followed by Norway with average APCs of 1817 euoros.

Table 1: top 20 producing publishers: publications numbers, APCs average, MNIJ and MNCS

Publisher	# publications	APCs average	MNIJ	MNCS
Springer Nature	14103	1 992 €	1,75	1,29
Elsevier BV	12534	2 855 €	1,99	1,99
Public Library of Science (PLoS)	9027	1 448 €	1,46	1,04
Wiley-Blackwell	6959	2 351 €	1,78	1,63
Frontiers Media SA	5725	1 686 €	1,25	0,95
MDPI AG	3438	1 212 €	1,16	0,85
Springer Science + Business Media	3313	1 536 €	1,45	1,23
Oxford University Press (OUP)	3022	2 411 €	2,33	1,97
American Chemical Society (ACS)	2299	2 627 €	3,48	1,81
IOP Publishing	2127	1 569 €	1,55	1,37
Copernicus GmbH	1994	1 492 €	1,96	1,38
Informa UK Limited	1911	1 390 €	0,87	1,47
BMJ	1604	2 089 €	0,97	1,76
Royal Society of Chemistry (RSC)	1131	1 629 €	2,49	1,22
Optical Society of America (OSA)	905	1 891 €	1,72	1,91
SAGE Publications	829	929 €	0,85	1,55
Institute of Electrical & Electronics Engineers (IEEE)	803	1 505 €	1,82	3,12
The Royal Society	774	1 926 €	1,80	1,48
Hindawi Publishing Corporation	739	1 370 €	0,67	0,53
Ovid Technologies (Wolters Kluwer Health)	715	3 215 €	1,41	1,82

Table 1 shows that more than half of publications are concentrated on the top 5 publishers. With a few exceptions, APCs are lower for large publishers than for the overall average. The MNIJ is higher than the world average for almost all publishers. Likewise for the MNCS. Furthermore, Table 1 shows that for some publishers, impact of publications (MNCS) is much higher than that of journals (MNIJ). This is particularly the case for the publishers "Informa UK Limited" and "BMJ". Explanation can be found in the fact that the majority of journals for these publishers (respectively, 87 and 93%) are either closed or hybrid. As demonstrated in the literature, OA publications are more cited than non-OA ones. Consequently, MNCS of publications indexed in OpenAPC database would be systematically higher than the average impact of journals in which they are published.

Table 2 shows that among the largest publishers, only Elsevier BV is listed in the top 20 most expensive ones (20th position), that are mostly American. We can also note that both impact of publications and impact of journals are high. Some exceptions can be made, especially for "MyJove Corporation" publisher with an average amount of 3081 euros for a very low impact. Similarly, "The American Association of Immunologists" charges for expensive APCs, while the average impact of journals and publications is at the level of the world average.

Table 2: top 20 most expensive publishers: publications numbers, APCs average, MNIJ and MNCS

Publisher	# publications	APCs average	MNIJ	MNCS
American Society for Nutrition	47	4 761 €	2,34	2,32
American Medical Association (AMA)	28	4 624 €	2,70	5,73
American Society of Clinical Oncology (ASCO)	23	4 588 €	1,35	2,63
Rockefeller University Press	66	4 466 €	3,64	2,32
American Psychological Association (APA)	132	3 754 €	1,45	1,86
American Society for Clinical Investigation	96	3 656 €	3,49	2,80
Royal College of Psychiatrists	64	3 630 €	1,45	1,98
EMBO	162	3 410 €	3,13	2,29
European Respiratory Society (ERS)	29	3 293 €	1,44	1,80
Ovid Technologies (Wolters Kluwer Health)	715	3 215 €	1,41	1,82
American Association for Cancer Research (AACR)	73	3 175 €	2,05	1,45
Nature Publishing Group	350	3 115 €	3,97	2,31
American Association for the Advancement of Science (AAAS)	166	3 085 €	4,97	3,36
MyJove Corporation	110	3 081 €	0,33	0,26
The Company of Biologists	414	3 017 €	1,76	1,08
The Endocrine Society	135	3 017 €	2,09	1,42
Society for Neuroscience	233	2 982 €	2,53	1,56
The American Association of Immunologists	87	2 976 €	1,08	1,06
Mary Ann Liebert Inc	117	2 872 €	1,21	1,08
Elsevier BV	12534	2 855 €	1,99	1,99

Correlation test

We performed a Spearman correlation test on four variables at publisher level: publications number (pub_nbr), APCs average (APCs_avg), MNIJ and MNCS. The results are presented in figure 4.

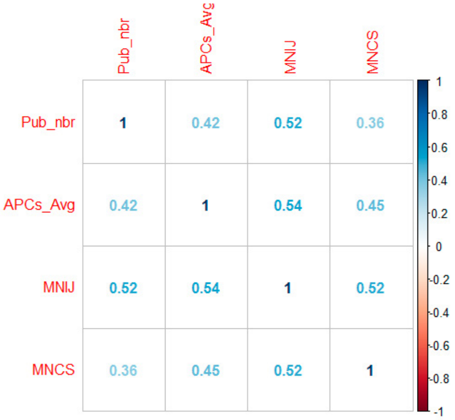


Figure 4: Spearman correlation matrix

All coefficients are significant at 5%. We observe that the number of publications is more correlated with the MNIJ (0.52) than with the MNCS (0.36). This means that within the OpenAPC database there are small publishers whose publications have high citations scores and large publishers with low citations scores. We also note that the number of publications per publisher is moderately correlated with the amount of APCs. This would mean that there is no evidence of relationship between the size of publisher and the amount of APCs.

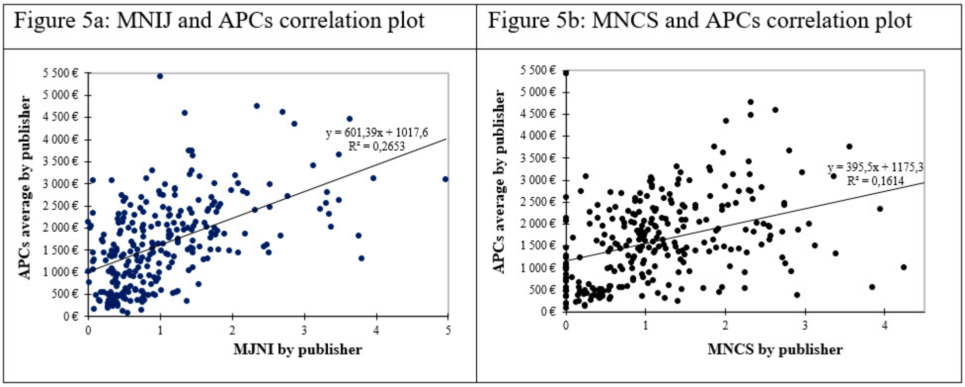


Figure 5a: MNIJ and APCs correlation plot

Figure 5b: MNCS and APCs correlation plot

Figure 5: APCs average per publication

Figure 3 shows correlation plots between on the one hand MNIJ and APCs (figure 5a), and on the other hand the MNCS and the APCs (figure 5b). The correlation between MNIJ and the amount of APCs is much higher (0.54 against 0.45). This shows that publishers take the quality into account when pricing their journals. However, this prices does not necessarily translate into impact. As long as the APCs are only moderately correlated with the MNCS.

Regression results

Table 3 summarizes Tobit regression results for explaining NCS by the amount of APCs. Regression was carried out in two stages. First, only the explanatory variable was integrated (*Log APC* per publication - model 1). Then, control variables were added (model 2).

Table 3: Tobit maximum likelihood estimation for log NCS

		Model_1		Model_2	
Variables_type	Variables	Coefficient	Pr(>|z|)	Coefficient	Pr(>|z|)
Explanatory	$Log\ (APCs)$	0.22***	$<2.22e^{-16}$	0.012***	$5.25e^{-05}$
Control	$Log(journal\_impact)$	-	-	0.471***	$<2e^{-16}$
	$Log(countries\_nbr)$	-	-	0.100***	$<2e^{-16}$
	Is_hybrid	-	-	0.204***	$<2e^{-16}$
Model statistics	Wald-statistic	2667	$<2.22e^{-16}$	$1.238e^{+04}$	$<2.22e^{-16}$
	Log-likelihood	$-8.063e^{+04}$	$<2.22e^{-16}$	$-7.607e^{+04}$	$<2.22e^{-16}$
	#publications	83,753			
	#Left-censored	11907			
	#Uncensored	71846			
	#Right-censored	0			

****Significant at 1%*

Table 3 shows that when control variables are not taken into account, the amount of APCs strongly impacts citation score (model 1). Once control variables are integrated, the APCs amount impact drops significantly. However, it can be seen that the amount of APCs has a positive impact on citations. Another interesting result is the impact of hybrid journals. Thus, if the journal is hybrid, citations score is higher. In other words, OA articles published in hybrid journals are generally more cited than OA articles published in 100% APCs journals. This is to be expected, given that not all well-established journals in the market have adopted a fully OA business model (Traag and Waltman, 2019). On the other hand, the main large publishers have massively integrated the hybrid model from 2013 (Besancenot and Vranceanu, 2017). In contrast, many fully OA journals are recently created journals that still have not built such a strong reputation for quality (some might even be aiming for average-level reputation and impact if this maximizes income – see (Traag and Waltman, 2019)). Hybrid journals are therefore more likely to be, at moment, in a virtuous circle where they receive higher quality manuscripts than fully OA journals, which translates to higher NCS of the published articles.

Conclusion and discussion

Through this article, we seek to analyze relationship between APCs and academic impact. Based on a large sample of 83,752 publications our study empirically verifies the belief that if we pay dearly for publication, impact of publication would necessarily be high. This belief stems from the fact that an author or an institution may think that all publishers who charge a high price for APCs and indexed in international databases like WoS, necessarily have a high academic quality. Our results show that contrary to this belief, paying dearly does not necessarily increase impact of publications. First, large publishers with high impact are not the most expensive in terms of APCs. Second, publishers with highest APCs are not necessarily the bests in terms of impact. Correlation between APCs and impact is moderate.

Otherwise, in the econometric analysis we have shown that publication quality is strongly determined by journal quality in which it is published. This result agrees with several studies which show it empirically (Waltman and Traag, 2017; Maddi, Larivière and Gingras, 2019). International collaboration also plays an important role in citations score. This result is also consistent with literature (Larivière *et al.*, 2015).

Another interesting result relates to the impact of hybrid journals versus 100% APCs journals. The regression results indicate that if the journal is hybrid, the NCS is stronger than if it is fully open. This result is consistent with the study of (Schönfelder, 2020) on the same database (OpenAPC) which showed that journal's impact and hybrid status are the most important factors for the level of APCs.

Our results have several implications for public policy and authors choices when it comes to submit their publications. First, the strong interest for OA had an immediate effect on the publishing market. Prices of OA publications have increased exponentially. This increase is disproportionate to the academic impact. The impact of publications for which authors have dearly paid is no better than that of publications with low APCs. Impact may even be lower. We also showed that some publishers are taking advantage of OA movement to demand high APCs, while their academic impact is very low. Finally, our results suggest that, for authors, APCs should not be used as an indicator for journals selection for submission. For institutions, for efficient management, it is important to be attentive to journals quality before granting funds for OA publication.

References

Asai, S. (2019) 'Determinants of Article Processing Charges for Medical Open Access Journals', Journal of Electronic Publishing, 22(1). doi: 10.3998/3336451.0022.103.

Besancenot, D. and Vranceanu, R. (2017) 'A model of scholarly publishing with hybrid academic journals', Theory and Decision, 82(1), pp. 131–150. doi: 10.1007/s11238-016-9553-0.

Björk, B.-C. (2004) 'Open Access to Scientific Publications – An Analysis of the Barriers to Change?' Available at: https://helda.helsinki.fi/dhanken/handle/10227/647 (Accessed: 6 March 2020).

Björk, B.-C. and Solomon, D. (2012) 'Open access versus subscription journals: a comparison of scientific impact', BMC Medicine, 10(1), p. 73. doi: 10.1186/1741-7015-10-73.

Björk, B.-C. and Solomon, D. (2015) 'Article processing charges in OA journals: relationship between price and quality', Scientometrics, 103(2), pp. 373–385. doi: 10.1007/s11192-015-1556-z.

Bruns, A., Rimmert, C. and Taubert, N. (2020) 'Who pays? Comparing cost sharing models for a Gold Open Access publication environment', arXiv:2002.12092 [cs]. Available at: http://arxiv.org/abs/2002.12092 (Accessed: 4 March 2020).

Cary, M. and Rockwell, T. (2020) 'International Collaboration in Open Access Publications: How Income Shapes International Collaboration', Publications, 8(1), p. 13. doi: 10.3390/publications8010013.

Chi Chang, C. (2006) 'Business models for open access journals publishing', Online Information Review, 30(6), pp. 699–713. doi: 10.1108/14684520610716171.

Copiello, S. (2020) 'Business as Usual with Article Processing Charges in the Transition towards OA Publishing: A Case Study Based on Elsevier', Publications, 8(1), p. 3. doi: 10.3390/publications8010003.

Ghane, M. R., Niazmand, M. R. and Sabet Sarvestani, A. (2020) 'The citation advantage for open access science journals with and without article processing charges', Journal of Information Science, 46(1), pp. 118–130. doi: 10.1177/0165551519837183.

Gumpenberger, C., Ovalle-Perandones, M.-A. and Gorraiz, J. (2013) 'On the impact of Gold Open Access journals', Scientometrics, 96(1), pp. 221–238. doi: 10.1007/s11192-012-0902-7.

Khoo, S. (2019) 'Article Processing Charge Hyperinflation and Price Insensitivity: An Open Access Sequel to the Serials Crisis', LIBER Quarterly, 29(1), pp. 1–18. doi: 10.18352/lq.10280.

Larivière, V. et al. (2015) 'Team size matters: Collaboration and scientific impact since 1900', Journal of the Association for Information Science and Technology, 66(7), pp. 1323–1332. doi: 10.1002/asi.23266.

Maddi, A. (2020) 'Measuring open access publications: a novel normalized open access indicator', Scientometrics, 124(1), pp. 379–398. doi: 10.1007/s11192-020-03470-0.

Maddi, A. and Gingras, Y. (2021) 'Gender Diversity in Research Teams and Citation Impact in Economics and Management', Journal of Economic Surveys, 0(0), pp. 1–24. doi: https://doi.org/10.1111/joes.12420.

Maddi, A., Lardreau, E. and Sapinho, D. (2021) 'Open access in Europe: a national and regional comparison', Scientometrics. doi: 10.1007/s11192-021-03887-1.

Maddi, A., Larivière, V. and Gingras, Y. (2019) 'Man-woman collaboration behaviors and scientific visibility: does gender affect the academic impact in economics and management?', 17th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS ISSI2019 with a Special STI Indicators Conference Track, 2(1687–1697), p. 12.

OST (2019) Dynamics of scientific production in the world, in Europe and in France, 2000-2016. Paris, France: HCERES, p. 100. Available at: https://www.hceres.fr/fr/publications/dynamics-scientific-production-world-europe-and-france-2000-2016-ost (Accessed: 23 September 2020).

Pinfield, S., Salter, J. and Bath, P. A. (2016) 'The "total cost of publication" in a hybrid open-access environment: Institutional approaches to funding journal article-processing charges in combination with subscriptions', Journal of the Association for Information Science and Technology, 67(7), pp. 1751–1766. doi: 10.1002/asi.23446.

Piwowar, H. et al. (2018) 'The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles', PeerJ, 6, p. e4375. doi: 10.7717/peerj.4375.

Prosser, D. (2003) 'The Next Information Revolution - How Open Access repositories and Journals will Transform Scholarly Communications', LIBER Quarterly, 14(1). doi: 10.18352/lq.7755.

Rizor, S. L. and Holley, R. P. (2014) 'Open Access Goals Revisited: How Green and Gold Open Access Are Meeting (or Not) Their Original Goals', Journal of Scholarly Publishing. doi: 10.3138/jsp.45.4.01.

Schönfelder, N. (2020) 'Article processing charges: Mirroring the citation impact or legacy of the subscription-based model?', Quantitative Science Studies, 1(1), pp. 6–27. doi: 10.1162/qss_a_00015.

Solomon, D. J. and Björk, B.-C. (2012) 'A study of open access journals using article processing charges', Journal of the American Society for Information Science and Technology, 63(8), pp. 1485–1495. doi: 10.1002/asi.22673.

Sotudeh, H. and Estakhr, Z. (2018) 'Sustainability of open access citation advantage: the case of Elsevier's author-pays hybrid open access journals', Scientometrics, 115(1), pp. 563–576. doi: 10.1007/s11192-018-2663-4.

Sotudeh, H., Ghasempour, Z. and Yaghtin, M. (2015) 'The citation advantage of author-pays model: the case of Springer and Elsevier OA journals', Scientometrics, 104(2), pp. 581–608. doi: 10.1007/s11192-015-1607-5.

Tananbaum, G. (2003) 'Of wolves and and boys: the scholarly communication crisis', Learned Publishing, 16(4), pp. 285–289. doi: 10.1087/095315103322422035.

Traag, V. and Waltman, L. (2019) 'Persistence of journal hierarchy in open access publishing', in 17th International conference on scientometrics and informetrics. STI2019, Rome, Italy, pp. 1339–1345. Available at: https://zenodo.org/record/3250081/files/483_Plan%20S.pdf?download=1.

Waltman, L. et al. (2011) 'Towards a new crown indicator: Some theoretical considerations', Journal of Informetrics, 5(1), pp. 37–47. doi: 10.1016/j.joi.2010.08.001.

Waltman, L. and Traag, V. A. (2017) 'Use of the journal impact factor for assessing individual articles need not be wrong', arXiv:1703.02334 [cs]. Available at: http://arxiv.org/abs/1703.02334 (Accessed: 6 March 2020).

Zhang, L. and Watson, E. M. (2017) 'Measuring the Impact of Gold and Green Open Access', The Journal of Academic Librarianship, 43(4), pp. 337–345. doi: 10.1016/j.acalib.2017.06.004.

In-Cites research fronts and its relationship with citations per document. A case study of Carlos III University of Madrid

Jorge Mañana-Rodríguez[1], Núria Bautista-Puig[2] and Elías Sanz-Casado[1]

[1] *jmanana@pa.uc3m.es, elias@bib.uc3m.es*
Research Institute for Higher Education and Science (INAECU), Carlos III University of Madrid, 126 Madrid Str., 28903, Getafe (Spain)

[2] *Nuria.Bautista.Puig@hig.se*
University of Gävle, Department of Industrial Management, Industrial Design and Mechanical Engineering, Kungsbäcksvägen 47, 801 76 Gävle (Sweden)

Abstract

Research fronts (RFs) represent the most dynamic areas of science and technology and the topics that receive higher attention in a specific field. Their identification has become the focus of global scientific and technological competition. In this study, we performed an analysis of the In-Cites research fronts in a case study for the scientific output of the Carlos III University of Madrid. This research in progress has two objectives: a) to quantify the alignment of the research output of UC3M in the different fields with the RF and b) to test if papers aligned with RFs receive more citations per document than those not aligned. The ultimate goal of this research is to assess the usefulness of RFs as an orientation for the establishment of research priorities at the institutional level for the university.

Introduction

Research Fronts (hereafter RFs) have been an increasingly relevant object of research in scientometrics. The literature on this topic grew steadily in the last century and has received wide attention over the last two decades (Szomszor, Pendlebury & Rogers, 2020). RFs have been labeled 'the footprint of the scientific communities' (Fajardo-Ortiz et al., 2017). RFs have become of interest for several audiences (Small, Boyack & Klavans, 2014). Nevertheless, there is no widely accepted definition detailing the specific attributes of research fronts and emerging research topics and technologies. As Rotolo, Hicks & Martin (2015, p.1827) point out 'the lack of consensus over definitions is matched by an eclectic and ad hoc approach to measurement'. Their definition of a research front includes, as Wang (2018) summarizes, five attributes: radical novelty, relatively fast growth, coherence, prominent impact and uncertainty and ambiguity. Wang's work proposes, develops and applies four criterions: growth (rapid increase in yearly publications), novelty (novel atthe early stage of its emergence), scientific impact (prominent volume of citations) and coherence (number of within-cluster citations divided by the total number of publications). A closely related topic to RFs, that of 'emergence' (in the case of emerging technologies) has been defined by Small, Boyack & Klavans (2014) as participating from two key features: novelty and growth. Some authors have underlined (Li & Chu, 2017) the relatedness between RFs and the dynamic and multi-dimensional conception of research activity.

Different studies have developed quantitative methods that can be used to identify and characterize RFs and their evolution over time. There is an extensive corpus of research using citation networks as a methodology for the identification of RFs. Mainly, as pointed by Shibata et al. (2009) research on the topic presents two approaches: one related to time-based indicators (i.e. future citations are predicted by current citations) and the other based on the detection of emerging clusters of densely connected papers. Both involve methods such as direct citation or bibliographic coupling. However, these approaches present some limitations: citation bias depending on authors' citing motivations (e.g. citation of their colleagues); it takes longer for fields with smaller scales or developing slowly (Shibata et al., 2008) and the fact that higher citation frequency does not necessarily imply greater quality (Marrone, 2020). In recent years, different approaches using text mining methods have been generalized. These include techniques such as Latent Class Analysis (LCA) or Latent Semantic Analysis (LSA) (Marrone, 2020). As an example, the burst term detection has become popular to explore dynamics and new trends (e.g. the Kleinberg burst detection algorithm). However, as pointed by Li & Chu (2017) these techniques are reliant on stochastic models which are particularly sensitive to the initial conditions. Another approach is the set of topic-modeling methodologies (Marrone, 2020). While these attempts allow discovering meaningful categories called 'topics' that can be used to understand how research fields change over time, such methods present some caveats such as the inconsistency in the terminology, which may reduce their usefulness or the ability to name unlabeled topics. Mapping and visualization techniques on the identification of the RFs (e.g. Chen, 2006) have also gained popularity in recent years. Previous studies identifying RFs were developed in a wide variety of fields such as engineering (Shibata et al., 2009), medicine or health-related research (Schwechheimer & Winterhager, 2001), technology (MacDonald & Dressler, 2018), or scientometrics (Anegón, Contreras & Corrochano, 1998), among others. However, to the best of our knowledge, not many approaches focused at an institutional level as we presented in this case study.

Clarivate Analytics counts with two products that provide information on RFs. On the one hand, InCites, which a customized research evaluation tool allowing institutional benchmarking on output, citation, collaboration, influential work and the discovery of areas of opportunity. Research Fronts are one of its components. On the other hand, Essential Science Indicators (ESI) provides data on Highly Cited Papers (articles in the top 1% of citations in the same field and year) and Hot papers (articles in the top 0.1% of citation frequency in the most recent bimonthly period for the same field and year). The filtering options include research fields, authors, institutions, countries and regions and research fronts. The RFs at the ESI interface can also be filtered by research fields and they are provided as a set of five strings separated by semicolon. These strings are the result of the "analysis of frequently occurring keywords or phrases in the titles of the paper" (Clarivate Analytics, 2021). In total, the interface provides 11,769 research fronts for 22 fields. We provide an example as an endnote[i]. The output also contains the number of highly cited papers, hot papers or both (named 'Top Papers' in the filtering options) in (it is to say, being the articles that conform the RF) the RF, cites to highly cited papers and cites per highly cited paper as well as the mean year. The identification and definition of the RFs (Clarivate Analytics, 2021) is based on single-link clustering of highly cited papers via normalized co-citation. Co-citation frequency is divided by the square root of the product of the citation frequencies of the two (initial) papers, aggregating other pairs that share common papers. The measure of association applied is the number of times pairs of papers have been co-cited.

The main attributes of the RF provided by Clarivate Analytics are high impact (since they are calculated on HCP) and currency (due to the yearly basis of the calculation and updating of the RFs). As Clarivate Analytics (2021) put it, one of the practical applications of RFs is to *"(...) assist in identifying areas where important work is being done and where the scientific community is focusing its attention*". This theoretical use of the RFs counts with its practical counterpart in their use for the identification of strengths and areas of opportunity for institutions in the product In-Cites, as well as in the reports produced in collaboration with the Chinese Academy of Sciences, in its seventh installment in 2021 (Clarivate Analytics and CAS, 2021).

Taking into account the prior practical uses of the RF our first objective is to quantify the alignment of the research output of UC3M in the different fields with the RF. Such alignment might be useful for decision making at the University concerning areas of attention since both, scientific impact and currency of research are desirable features of the research output of the University. For the above rationale to be of practical use we need to prove that, beyond HCP and for the 'regular' output of the university in the various fields (on which the RFs are defined), there is a positive relationship between citations and alignment. This might seem plausible at first, but the association between a greater scientific impact and the alignment with the RF, albeit obvious for the HCP, could take a diversity of values for non-HCP papers in various fields or even not be present at all, thus requiring empirical contrast. For this reason, the second objective of this research is to analyze if there is an association (causal or otherwise) between the alignment with RFs and citations for a set of articles from UC3M that are not HCP.

Data and methods

Dataset

The workflow used in this study is summarized in Figure 1. We downloaded all the registers (articles only) of the UC3M at the Web of Science's Core Collection for the years 2015-2019. For this purpose, we used the OG (organization enhanced) field and retrieved only the articles included at the Social Sciences Citation Index (SSCI), Science Citation Index (SCI) and Arts and Humanities Citation Index (A&HCI). We selected these databases because they encompass most of the publications registered in the three major branches of knowledge. In parallel, we collected all the literal descriptors of the RFs available at the 2020 Essential Science Indicators interface, by research area as in the case of the example provided in the introductory section.

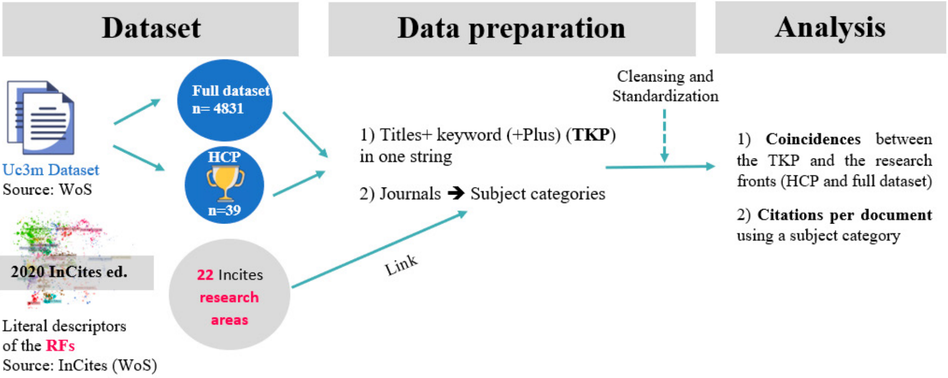


Figure 1. Workflow used in this study

Data preparation

1) In this step, we merged the titles, keywords and keywords plus (TKP) into a single string for each article. We also removed from the matching process all the coincidences of single terms *(Cleansing and Standarization),* given the greater likelihood that these terms are void or spurious in terms of their relationship to the strings defining each research front. This was intended to provide a more conservative and robust measure of the alignment of the documents with the research fronts.

2) We linked each subject category in which the journals of our dataset are classified to one of the 22 In-Cites research areas (not all WoS journals are classified in an ESI field). For this purpose, we used the list of journals in each In-Cites category and identified the frequency of their WoS category, using the maximum frequency of coincidence as rule of attribution of a WoS category to an In-Cites category. If an article's journal is classified in two subject categories that belong to the same In-Cites classification, the article was counted as one whole publication classified in that single research area. In the case of a journal classified in *n* subject categories that were mapped to different In-Cites categories, we calculated a direct fractional score using the formula 1/n, where n is the number of different research areas associated to the subject categories. We then summed the fractional scores for each of the 22 research areas. That value is the number of contributions per research area.

Analysis

The analysis is conducted in two steps.
1) In a first step we calculated the number of coincidences between the TKP and the RFs. We searched, for each RF definition the coincidence of terms from the contributions' merged titles, keywords and keywords plus in the same field. In example, we found that an article in the area of Economics contained the string 'corporate social responsibility', which is one of the four terms of one of the research fronts in that field. The result of this process was the number of coincidences between each article's TKP and the RF in their own research area for the overall output of the UC3M-
2) We also calculated the citations per document using the fractional counting method. In a final step, we compared the citations per document using an In-Cites category scheme between the documents with at least one coincidence with the associated research area's RF and those without coincidences. We tested the existence of statistically significant differences in their median citations using Mann-Whitney's U test.

Results

Table 1 illustrates the median citations of the articles with and without coincidences with the RFs by In-Cites discipline. In all cases but Neuroscience & Behavior the median citations is greater in the case of articles with RF coincidences and in five of them these are statistically significant (p-value=.05).

Discussion and conclusions

The results do not allow establishing a causal relationship between RF alignment and citation potential, but yes a positive association. The results evidence a greater impact of RF-aligned research over non RF- aligned papers: this does not immediately follows from the definition of the RF (which are extracted only from HCP), but rather suggest the existence of a continuum of alignment between papers and RFs beyond the boundaries of the HCP set. If this continuum exists, the 'distance' or 'alignment' with RF seems to be non-independent from the existing citation count and possibly from the citation potential of scientific literature. Although the data is not fully developed yet, preliminary comparisons of citations per document by field of knowledge show that UC3M output in several fields is below the median for the public Spanish University System. This might also correspond with a lower alignment of its research lines with the RFs in the case of UC3M. Despite the lack of proven causality, the clear association between alignment and impact underlines the usefulness of taking the descriptors of the RFs as potential points of reference for decision-making concerning the prioritization of research lines, resource allocation, and economic incentives (Chen and Guan 2011).

Table 1. Descriptive statistics for the contrast of medians between the articles aligned and not aligned with the RF at UC3M.

	N. Articles with coincidence with RF	*N. Articles without coincidence RF*	*% aligned with RF*	*Median citations (coincidence)*	*Median citations (no coincidence)*	*Bilat. Sig.*
Engineering	573	951	37.60	2.5	2	0.003
Computer Science	211	454	31.73	2.33	1.5	0.001
Physics	172	882	16.32	3	3	0.707
Mathematics	146	503	22.50	2	1.33	0.004
Materials Science	131	280	31.87	3	2.5	0.168
Chemistry	129	330	28.10	3	2	0.003
Social Sciences, General	111	283	28.17	2	1	0.023
Economics & Business	94	314	23.04	3.5	2	0.001
Clinical Medicine	54	46	54.00	3	2.17	0.219
Neuroscience & Behavior	27	20	57.45	3.5	4.08	0.647
Environment/Ecology	22	67	24.72	3.67	2.5	0.066
Psychiatry/Psychology	15	27	35.71	5	2	0.168

Note. Biology & Biochemistry, Geosciences, Pharmacology, Agricultural Sciences, Immunology, Space Sciences, Plant & Animal Science and Molecular Biology and Genetics count with 7 or less articles with coincidences. These values are considered too low to allow extracting useful conclusions and for that reason are not included in the table.

This research presents some limitations that the authors intend to overcome in the near future. HCP considered in this study are not limited to a given time period, which might imply a form of bias in the analysis of the RFs alignment (an HCP from 2012 might present spurious alignment with current, 2019 RF). In addition, these results can be considered valid for the UC3M outputs, but further research implies the necessity of testing the congruence of these results with other institutions of different types (technical, generalist, established and new universities). It has not been possible to establish a causal relationship between citations and alignment with RFs and this might limit the usefulness of the conclusions for the university, albeit RFs are systematically used for similar objectives in Clarivate Analytics' In-Cites. In this sense, despite already taking into account some of the main sources of variability in citations (the same institution and most lecturers and their output divided by fields of knowledge), we expect to isolate other intervening sources of variability maximizing the *ceteris paribus* scenario in further research. Concerning the methodology used for counting the coincidences, the authors acknowledge that more sophisticated procedures have been developed and tested for text-based relatedness measures, such as those detailed in section 3.2 of Waltman et al (2020). In this first approach, we used a simpler methodology in terms of co-occurrence in order to serve as a proof of concept, but further developments would imply the use of the aforementioned measures and procedures when applicable. Finally, we only counted coincidences as a dichotomous variable: an article presents a coincidence with a RF if two or more terms coincide. The degree of alignment of an article with a RF could vary substantially taking into account the number of coincidences for a single paper (i.e. paper 'a' presents 1 coincidence whereas paper 'b' presents four coincidences', both with the same RF). That gradation ought also to be addressed in further research.

References

Anegón, F. D. M., Contreras, E. J., & Corrochano, M. D. L. M. (1998). Research fronts in library and information science in Spain (1985–1994). *Scientometrics*, *42*(2), 229-246.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, *57*(3), 359-377.

Chen, K., Guan, J. (2011). A bibliometric investigation of research performance in emerging nanobiopharmaceuticals. *Journal of Informetrics*, 5(2), 233-247.

Clarivate Analytics (2021). Essential Science Indicators Help (http://esi.help.clarivate.com/Content/research-fronts.htm)

Clarivate Analytics and CAS (2021. Research Fronts. Active Fields, Leading Countries. *Clarivate Analtics and Chinese Academy of Sciences*. Available under demand at: https://clarivate.com/webofsciencegroup/article/seventh-annual-research-fronts-report-highlights-hot-and-emerging-fields/

Fajardo-Ortiz, D., Lopez-Cervantes, M., Duran, L., Dumontier, M., Lara, M., Ochoa, H., & Castano, V. M. (2017). The emergence and evolution of the research fronts in HIV/AIDS research. *PLoS ONE*, *12*(5), 1–13. https://doi.org/10.1371/journal.pone.0178293

Li, M., & Chu, Y. (2017). Explore the research front of a specific research theme based on a novel technique of enhanced co-word analysis. *Journal of Information Science*, *43*(6), 725–741. https://doi.org/10.1177/0165551516661914

MacDonald, K. I., & Dressler, V. (2018). Using citation analysis to identify research fronts: A case study with the internet of things. *Science & Technology Libraries*, *37*(2), 171-186.

Marrone, M. (2020). Application of entity linking to identify research fronts and trends. *Scientometrics,* 122(1), 357–379. https://doi.org/10.1007/s11192-019-03274-x

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology?. *Research policy*, *44*(10), 1827-1843.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. Research Policy, 43(8), 1450–1467.

Schwechheimer, H., & Winterhager, M. (2001). Mapping interdisciplinary research fronts in neuroscience: A bibliometric view to retrograde amnesia. *Scientometrics*, *51*(1), 311-318.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, *28*(11), 758-775.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for information Science and Technology*, *60*(3), 571-580.

Szomszor, M., Pendlebury, D., & Rogers, G. (2020). Global Research Report - Identifying Research Fronts in the Web of Science : From metrics to meaning. *Web Of Science*, *September*, 1–20. https://clarivate.com/webofsciencegroup/article/identifying-research-fronts-in-the-web-of-science-from-metrics-to-meaning/

Wang, Q. (2018). A bibliometric model for identifying emerging research topics. *Journal of the association for information science and technology*, *69*(2), 290-304.

Waltman, L., Boyack, K. W., Colavizza, G., & van Eck, N. J. (2020). A principled methodology for comparing relatedness measures for clustering publications. *Quantitative Science Studies*, *1*(2), 691-713.

Endnotes

[i] One of the research fronts in mathematics is defined by the following set of strings: CONSTRUCTING LUMP SOLUTIONS; LUMP KINK SOLUTIONS; LUMP SOLUTIONS; GENERALIZED (2+1)-DIMENSIONAL HIROTA BILINEAR EQUATION; INTERACTION SOLUTIONS

A comparison of citation scores and Web of Science usage counts for Russian science papers, 2006-19

Valentina Markusova [1], Grant Lewison [2], Anna Zolotova [1], Ilya Libkind [3] and Alexander Libkind [1]

[1] *valentina.markusova@gmail.com, korablik@mail.ru, anliberty@mail.ru*
Russian Institute for Scientific and Technical Information (VINITI) of the RAS, Moscow 125119, (Russia)
Institute for the Study of Science of the RAS, Moscow, 117218, (Russia)

[2] *grantlewison@aol.co.uk*
King's College London, Department of Cancer Policy, Guy's Hospital, Great Maze Pond, London SE1 9RT, (UK)

[3] *libkind_ilya@hotmail.com*
The Russian Federation Financial University, Leningradsky Prospect 49, Moscow 125119, (Russia)

Abstract

We examined research papers from the Russian Federation during the 14 years, 2006-19, in chemistry, physics, mathematics, and biomedicine, that were covered in the Web of Science (WoS) Science Citation Index – Extended. We saw a modest expansion in physics, chemistry, and mathematics which are traditional Russian strengths. However, following directives from VV Putin, biomedical research is now also being encouraged. We compared the mean numbers of Total Citations (TC) to papers from different years with their usage counts in the WoS. U1 shows the numbers of usages in the last 180 days, and U2 gives the count since 2013. These "altmetrics" provide an alternative means for research evaluation, and can give an indication of the papers' interest and utility to other researchers earlier than citation counts. Usage counts (U2) are mostly higher than TC for recent years. Russian physics papers are the most cited, but chemistry papers receive more usage counts. Biomedical research has expanded, more than the other subject areas of our study, but it is still not well cited.

Introduction

Traditional scientometrics, as originally defined (Nalimov & Mulchenko, 1969) meant "the quantitative methods of research on the development of science as an informational process". It involves the counting of papers, and citations to them from other papers, and their classification in various ways. The production of the world in particular subject areas, of individual countries, of research institutions (such as universities), of research groups, and even of individual scientists, can all be counted, and compared with those of others in order to assist the process of research evaluation. More recently, with the rise of social media and many other types of document, the concept of citations has been widened to include these citations, which are often referred to as "altmetrics". Altmetrics is "the study and use of scholarly impact measures based on activity in online tools and environments" (Priem *et al*, 2012). Also called Scientometrics 2.0, this field complements journal citations with impacts in social networks such as views, downloads, "likes," blogs, Twitter, Mendeley, CiteULike (Mingels *et al*, 2015). Research by Thelwall *et al* (2016) focused on medical publications (332,975 articles in 45 fields) indexed by Scopus in 2009. This study demonstrated a strong correlation between citations and Mendeley readership counts across all medical fields. According to the authors' opinions, altmetric indicators should be used as a substitute for classical metrics. The field has spawned many learned articles, and even a new journal, the *Journal of Altmetrics Research*, which was founded by the late Judit Bar-Ilan in 2019.

In fact, there are two related fields, *webometrics* and *altmetrics*. The former is used to refer to references and other parameters that are found on the World Wide Web. These new research areas have provided some interesting insights into the world of scientific publishing, but they suffer from the disadvantage (from a scientific point of view) that they are essentially transitory, and so their findings cannot easily be repeated at a later date, which is a core requirement of scientific research. A significant growth of altmetric research, based on Web of Science data (WoS), was observed during the previous three years (Kousha *et al*, 2020, Fang *et al*, 2020). The peculiarity of classical citations metrics and altmetrics indicators were investigated in comprehensive bibliometric analysis conducted by Chi *et al* (2019) on two subfields of chemistry: analytical chemistry (C1) and organic chemistry (C3). Three types of indicators (Capture, Citation, and Usage Counts) were selected and extracted from the following platforms: WoS, Scopus, Mendeley, EBSCO, and CrossRef. The dataset of 39,736 papers in analytical chemistry (C1) and 27,531 papers in organic chemistry (C3) indexed by the WoS in 2013 was investigated. DOI identifications were used to match the dataset with PlumX. The results of the study showed significant correlations at different levels among various altmetric indicators. A high correlation was observed among classical citation indicators on the platforms that were studied.

However, not all these new indicators suffer from this lack of reproducibility. There are, in fact, several good alternatives to the counting of citations in the serial literature. One of the first to be exploited was the reference section of patents. These were found to be an indication of the scientific underpinning of the inventions, and they included not only other prior patents, but also scientific papers from journals and conference proceedings. It was found that if a patent had many non-patent references, it was likely to be valuable (Harhoff *et al.*, 1999). Another finding was that the research cited in this way was often basic rather than applied (Narin & Olivastro, 1998). Another form of citation that has been examined for some years, but has not been widely used because there is no convenient database, is the way the mass media, and in particular newspapers, publish stories about research. These are particularly frequent in medical research, and allow scientists' discoveries to be appreciated by the general public, and also by policy makers, and other scientists, who may give more academic citations to the cited papers as a result of this publicity (Phillips *et al*, 1991; Lewison *et al.,* 2008).

For the last two decades (Grant, 1999; Grant *et al*., 2000), the references on clinical practice guidelines have been used to show how medical research can be translated into clinical benefit through the recommendations that they provide, and which will increasingly be used to guide, and sometimes to determine, medical care. The references are usually chosen as a result of a careful search of the literature, and many potentially eligible papers are discarded because their methodology is not regarded as sufficiently rigorous. Recently, the process of collecting these references has been automated by Minso Solutions AB in Boras, Sweden (Eriksson *et al*., 2020; Pallari *et al*, 2021) so that a research funder, or a performer, can see which of their papers have been cited in this way. This process tends to strongly favour clinical work rather than basic research; it thus can help to redress the balance because counts of citations in journal articles usually benefit the latter (van Eck *et al*., 2013).

In September 2015, the Web of Science (WoS, © Clarivate Analytics) introduced two new indicators of research impact. They are designated as U1 and U2. U1 counts the numbers of "usages" of a paper in the last 180 days, and U2 counts the numbers since 2013. "Usage" means that the person who has searched for a paper in the WoS has either attempted to download the full text, or has saved its record in a bibliographic reference manager (such as EndNote) or some other downloadable format. This is a new field of altmetrics, and it has led to several investigations that have attempted to compare these usage measures with citation counts (Markusova *et al*., 2018; Chi & Glänzel, 2019; Lewison, 2020). The results have been somewhat inconclusive because they are measuring different behaviours. In principle, a researcher would first identify a paper of interest by way of its title, then read the abstract (normally easily available), perhaps obtain a full text, especially if it is Open Access (OA), and then finally read the paper and cite it in a subsequent research paper of her own.

In this paper, we have examined Russian research production in four major fields: chemistry, mathematics and physics, the traditional areas of strength in Russian science, and one to which increased attention is now being given, namely biomedicine. Evaluation of research is now playing a greater role in the Russian Federation. Two presidential Decrees have concerned research. The first, #599 in May 2012, set a goal "that the Russian share of research productivity (RO) has to reach 2.44% of the global RO, and five Russian universities have to be among the top hundred universities included in one of three world ranking systems in 2015". There are further ambitious goals established in the Decree N 204 published in December 2018 (http://static.kremlin.ru/media/acts/files/0001201805070038.pdf). These would use research to propel the Russian economy into the top five of the world by 2024. The decree also established seven strategic priorities, and for the first time (Starodubov *et al.*, 2019) one of them was medicine. We will use different indicators to see if the Russian research system is responding to these challenges, and in particular, data on Total Citations (TC) and the two usage counts (U1 and U2) for Russian papers published during 2006-19.

Methodology

We downloaded the bibliographic details of all papers in the WoS with an address in Russia in the years 2006-19 in six subject categories (SC) by WoS classification: mathematics, 20 sub-fields of biomedicine, and two major sub-fields in each of chemistry and physics. These were inorganic & nuclear and physical chemistry, nuclear physics, and particles & fields. [The search was restricted to the Science Citation Index - Expanded (SCI-E). Because of the history of the science structure in Russia, which is focused heavily on hard science, there are only a small number of Russian social sciences and arts & humanities journals compared with natural sciences that are indexed in the WoS.] We then determined the number of papers each year in these six subject areas, and for each paper, the total number of citations to the date of retrieval in July 2020 (TC), and the values of U1 and U2 (see above). In total, we compiled data on 485,505 papers, which formed our database.

The main indicators for analysis were:

- The number of publications;
- The number and percentage of records that had at least one U1 value >= 1, one U2 value >= 1; and received at least one citation (TC)
- The mean number of citations (TC) per year;
- The mean number of U1 and U2 counts per year;
- The number and percentage of highly cited records;
- The percentage of OA publications in 2013. (Our hypothesis was that access to full-text publications could have an impact on usage counts.)

By way of example, mean values of TC, U1 and U2 were calculated as the quotient of total citations and usage counts to July 2020 (when the papers were identified and their details downloaded to file) divided by the number of papers published in a given year.

In order to see whether the research evaluation targets set by President V.V. Putin were met, or approached, we also determined the numbers of all papers in the SCI-E, and those in chemistry, mathematics and physics, and those in the defined subfields of biomedicine. These could then be compared with the respective totals with an address in the Russian Federation.

We made a detailed examination of the distribution of TC and U2 usage counts for the 2014 Russian papers in biomedical research (n = 6672). These papers were divided into individual groups with the same numbers of citations, or U2 counts, 0, 1, 2,3... Then the sum of citations and U2 counts for each group was calculated. These sums, as percentages of the respective totals, were then plotted on the ordinate axis of a graph, the abscissa being TC or U2. Not all values of TC or U2 were represented, and the gap between values, unity for small ones, became much larger for high values.

Results

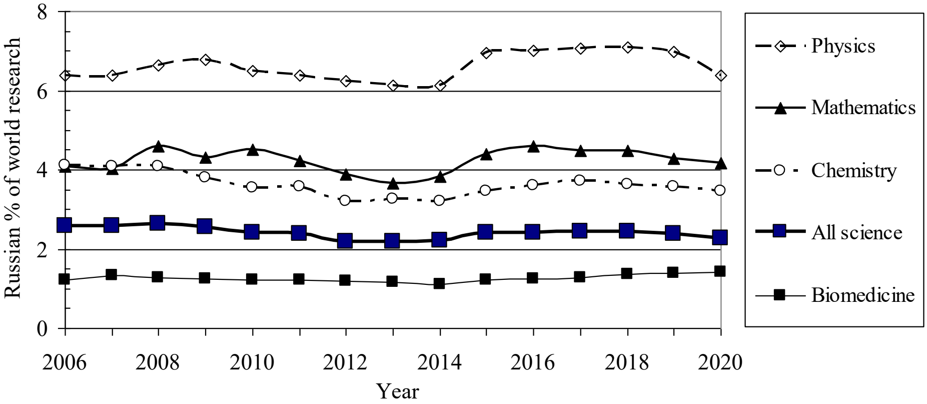


Figure 1. Russian research productivity in the Science Citation Index - Extended, 2006-20, in all science and in four major fields (integer counts), percentages of the world totals.

The first analysis was of Russian production in the four major fields, plus all science, and the comparison with world production is shown in Figure 1. Here Russian production is given as integer counts, *i.e*., full credit is given for each paper that contains at least one Russian address. It is clear that the target of Russia publishing 2.44% of world science was met easily for the three traditionally strong fields but not in biomedicine, where the average was only 1.26 %, though it did increase to 1.41% in 2020. There was a small but definite increase in all the percentages in 2015, mostly following a gradual decline from about 2008. But it has been followed by further small declines in the last two years. The Russian presence in all science has hovered around the target figure of 2.44%. It was slightly higher in 2006-09, but has since declined and dropped to 2.26% in 2020.

Among the 161 Russian journals indexed by SCI-E there are 41 journals in physics. Historically these journals are of high quality, six of them belonging to the top two citation quartiles (JCR, 2019). Russian science is moving rapidly towards open access (OA), which will improve its likelihood of being cited. Only 9% of its papers in 2006 were OA, but 15.4% were in 2013, and 28% were in 2020. However, the corresponding figures for the world were 20% in 2006 and 40% in 2020, because many of the major research funders in the USA and Europe are insisting that publications with their support should be OA, and are providing money for author publication fees (anon, 2008; Widener, 2018).

We turn now to the effects of Russian research in the six subject areas, two each in chemistry and physics, mathematics and biomedicine. The graphs, Figures 2 to 7, show the mean values of total citations (TC), and the two usage measures, U1 and U2. [The values of U1 are quite small, so 10 x U1 has been plotted so as to make them more visible, and any variations with time more obvious.] In all six graphs, mean U2 reaches a peak for papers published in 2013, because papers published earlier have "lost" some of their usage counts from their early years, when they normally attract the most interest. In the subsequent years, the curves of mean values of TC and U2 parallel each other. For most of the years, values of U2 exceed those of TC, indicating that scientists download more papers than they cite, as would be expected.

However, in mathematics and in physics: particles & fields, the two curves are almost superimposed, suggesting that for these fields a sight tends to lead to a cite. The lack of cites following a download is greatest for the two chemistry subject areas, where the ratio of U2 to TC is more than two for papers from 2012, and increases to four for 2016-17 papers, and six for the ones published in 2019. However, this may simply indicate that chemistry citations are much slower to appear than they are in other fields of science.

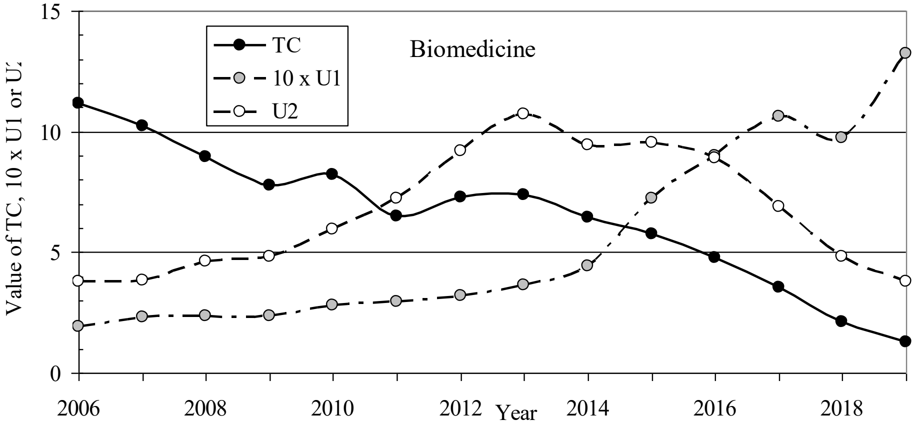


Figure 2. Mean values of total citations (TC) and usage counts U1 (multiplied by 10) and U2 for Russian biomedical papers in the WoS published in different years, 2006 to 2019.

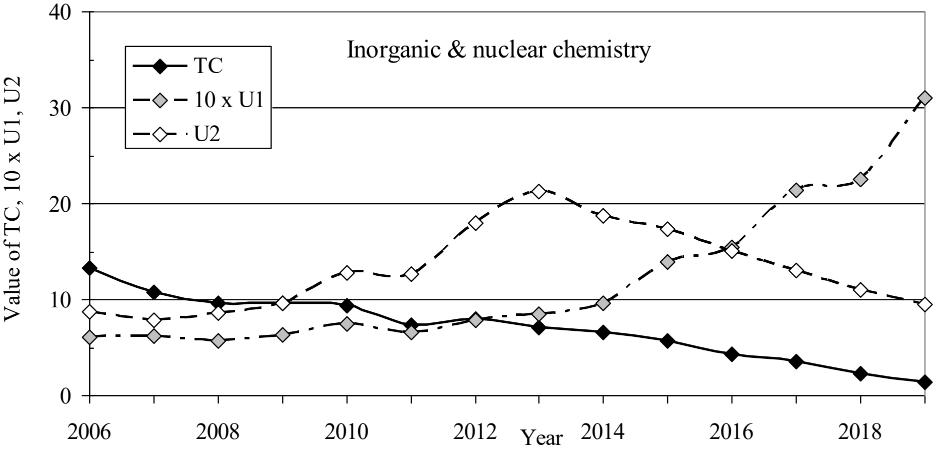


Figure 3. Mean values of total citations (TC) and usage counts U1 (multiplied by 10) and U2 for Russian inorganic & nuclear chemistry papers in the WoS published in different years, 2006 to 2019.

The values of U1 rise steadily for the most recent papers in biomedicine and both subject areas in chemistry, but not for mathematics and physics, where the peak occurs for 2017 papers. This indicates that in these subjects the maximum usage does not occur for papers in their first year, but takes two to three years to build up, as happens for citations. Chemistry papers receive a lot of usage in their first year, much more than those in physics (means are between one and two) and especially in mathematics (mean value below 0.3). This confirms what is known about this subject, that papers take a long time to emerge (some require many months to be checked by reviewers), and take even longer to have an impact.

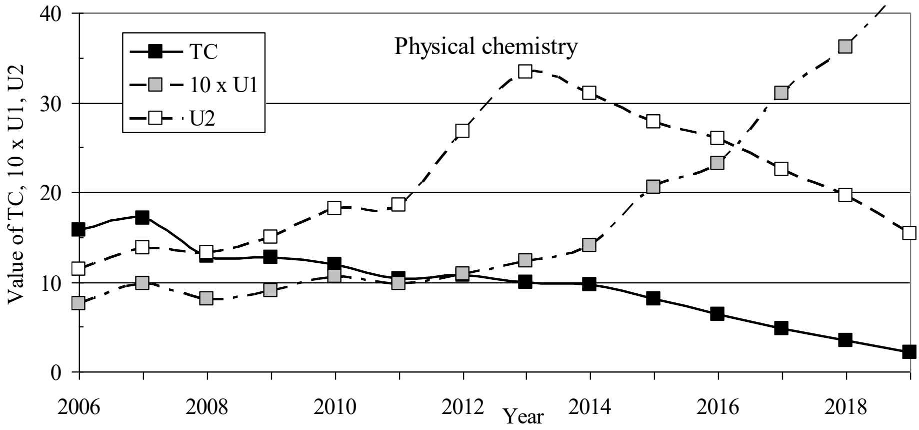


Figure 4. Mean values of total citations (TC) and usage counts U1 (multiplied by 10) and U2 for Russian physical chemistry papers in the WoS published in different years, 2006 to 2019.

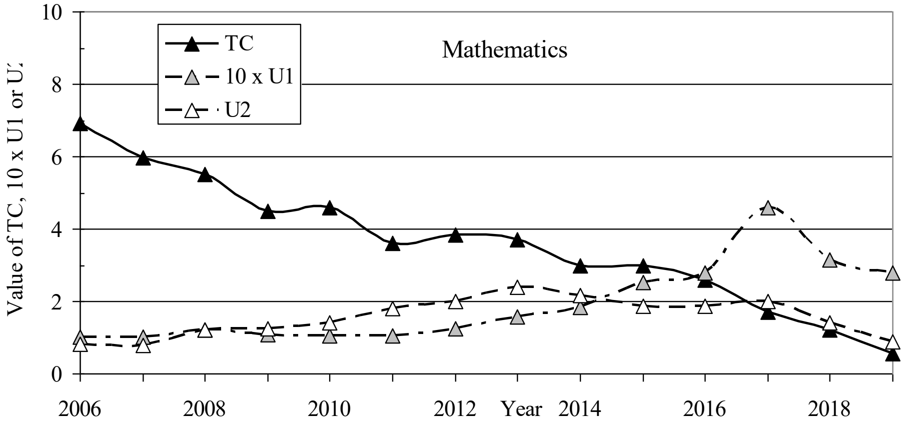


Figure 5. Mean values of total citations (TC) and usage counts U1 (multiplied by 10) and U2 for Russian mathematics papers in the WoS published in different years, 2006 to 2019.

It is very clear from these graphs that the most cited papers of these six groups were those in the two subject areas of physics. This held true for papers in all publication years, as Table 1 shows. There are several possible reasons for this, apart from the generally accepted high standing of the subject in the Russian Federation. One is that there is a high level of international collaboration: 59% of physics papers are collaborative compared with 32% for all Russian research productivity (RP). A consequence is that physics has a much higher share of highly-cited articles (ones in the top 1%), namely 1.12%, compared with just 0.66% for our dataset (InCites, 2015-19).

Table 1. Mean total citation (TC) scores for Russian papers in six subject areas in four years.

Year	Phys, P&F	Phys, Nucl	Chem, Phys	Chem, I&N	Biomed	Maths
2006	33.6	42.2	15.8	13.3	11.2	6.9
2010	22.0	14.4	11.9	9.4	8.2	4.6
2014	15.5	13.3	9.6	6.6	6.5	3.0
2018	8.2	4.3	3.5	2.2	2.1	1.2

Table 2. Mean usage count (U2) for Russian papers in six subject areas in four years.

Year	Phys, P&F	Phys, Nucl	Chem, Phys	Chem, I&N	Biomed	Maths
2006	3.1	6.1	11.4	8.7	3.8	0.8
2010	5.2	5.8	18.2	12.8	6.0	1.4
2014	10.0	13.1	31.0	18.8	9.4	2.1
2018	8.9	7.2	19.7	11.1	4.8	1.4

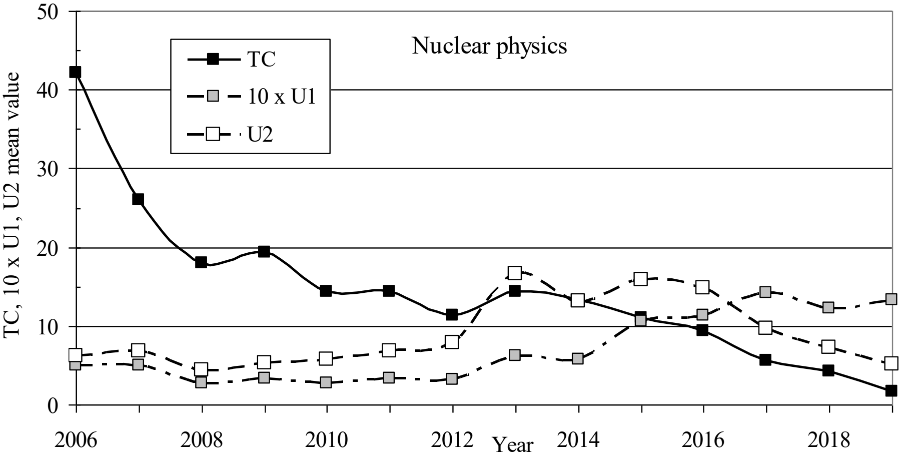


Figure 6. Mean values of total citations (TC) and usage counts U1 (multiplied by 10) and U2 for Russian nuclear physics papers in the WoS published in different years, 2006 to 2019.

On the other hand, the mean usage counts for the physics papers tend to be lower than for the chemistry papers, as seen in Table 2. This might be owing to an archive on high energy physics that has been operating since 1991 and is very popular among specialists (https://arxiv.org/archive/hep-th) (Chumachenko *et al.*, 2020). It is also noteworthy that in 2013, 42% of Russian particles & fields papers were published in OA journals, compared with the average of 15% for all Russian science publications. As mentioned above, the users of physics publications do not show a significant interest in very recent literature compared with users of research in other disciplines because the values of U1 peak in 2017 rather than in 2019. The height of the peak in U1 in 2017 for particles & fields (Figure 7) is 4.4, which is much the highest for the six disciplines studied here.

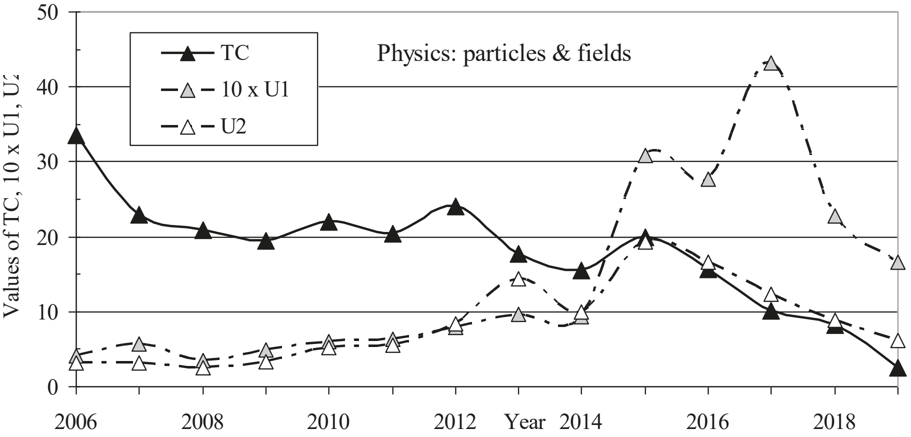


Figure 7. Mean values of total citations (TC) and usage counts U1 (multiplied by 10) and U2 for Russian physics: particles & fields papers in the WoS published in different years, 2006 to 2019.

The highest values of U2, in the peak year of 2013, are much higher for the two chemistry subject areas than for the other four, see Table 2. Researchers would have had to depend on abstracts, or to search for full-texts in the National Library because the OA percentages were only 3.6% in physical chemistry and 5.1% in inorganic & nuclear chemistry, far below the corresponding values for nuclear physics (25%) and particles & fields (42%). And a check on 12 chemistry papers with the highest U2 counts showed that they were closed access, with the full text only available through the journal publisher.

Mathematics has also been seen as a great Russian scientific strength, and Figure 1 shows that the Russian presence in the field is almost twice that of its presence in world science. However, its average TC score is much lower than that of the other five disciplines, see Table 1. The same is even more so with respect to U2, and this is perhaps surprising because the share of OA journals was relatively high at 23% in 2013.

As mentioned above, one of the strategic priorities selected by the Russian government is speedy development of “high-tech medical care and personalized medicine and health-saving technology”. The current pandemic showed the necessity of improvements in social medicine improvement in every country. The speedy development of the Russian vaccine “Sputnik-V” aroused great attention around the world. The creation of this vaccine is a tribute to many Russian scholars’ contributions to virology. It followed the development of an oral polio vaccine by Soviet microbiologists Professors K.M. Chumakov and M.I. Voroshilova. This saved the lives and health of millions of children in Africa in the 1960s (Kramer, 2020).

Biomedical research, although still only a small part of the overall Russian research portfolio, has been relatively successful in responding to this directive. The Russian percentage presence in the world has risen over the 14-year study period from 1.23 to 1.41%, or by 8%, see Figure 1. This may seem modest, but it contrasts with an increase of only 3% in physics, and *contractions* of 11% in all science and 14% in chemistry. Russian biomedical research accounted for 15% of all its scientific production in the WoS in 2019, and 23% of our database of Russian papers. However, it is still not well cited compared with the more traditional strong subjects, see Table 1. A consequence is that the leading Russian biomedical journals had very low citation Impact Factors: *Kardiologie* (0.13) and *Gematologiya i Transfiziologita* (0.06) in 2019. A recent study of cancer research in Eastern Europe (Begum *et al.*, 2018) revealed that the papers from Russia received fewer five-year citations on average than those of many other East European countries such as the Czech Republic, Estonia, Hungary, Poland, Slovakia, Slovenia and Ukraine.

The plot of the sums of TC and U2 values for the 6,672 biomedicine papers is shown in Figure 8. It shows that the two indicators are closely related. Only 14% of Russian publications in 2014 were OA, compared with a world average of 37% in that year.

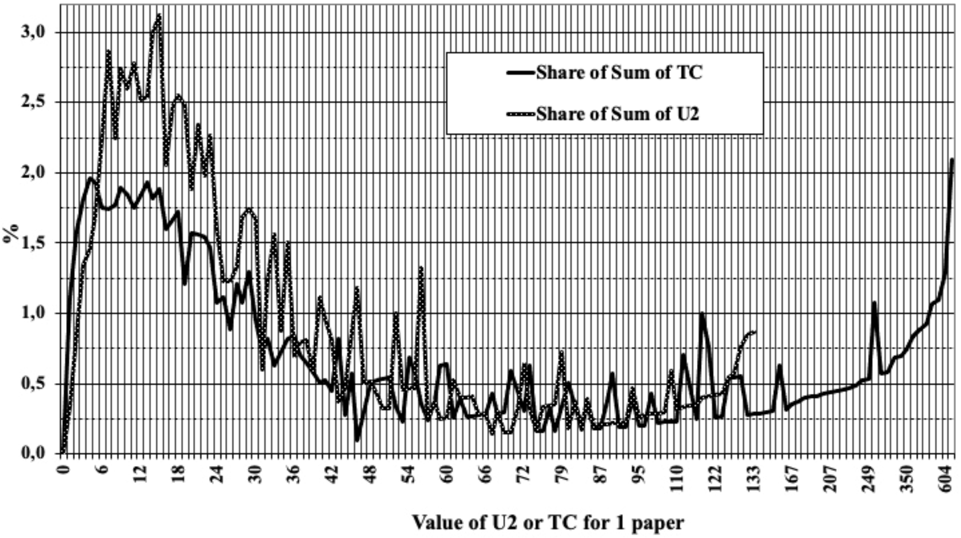


Figure 8. Share of sum groups with the same TC or U2 value for Russian biomedical research papers in 2014 in the SCI-E of the Web of Science.

Discussion

Our analysis of total citations (TC) and of counts of usage since 2013 (U2) showed a similar trend in both with time for all the subject areas that we studied except for mathematics. For the biomedical papers published in 2014, relatively few (15%) were effectively ignored by WoS readers and had U2 = 0, but many more (36%) were not cited between 2014 and 2020. It is notable (Fig. 8) that the TC distribution has risen because of the highly-cited articles (HCAs). In the whole dataset for 2014 there were 203 HCAs, and 28 of them were on biomedicine. The HCAs had a big influence (22%) on the total number of biomedical citations, but the highly-used articles (those among top 1%) made only a small contribution to the usage total (6.7%). Among these papers only one was published with domestic collaboration. It was in the Russian journal *Acta Naturae* (with an Impact Factor of only 1.36) which has been assigned to the Q4 journal group in the subject category Cell Biology. One article in particular (Makarov *et al*., 2014) received 484 citations (to July 2020), but only 82 usage counts. On the other hand, another one, the review article (Surmenev *et al.,* 2014) had the most usage counts (486) and also had many citations (389). Most (78%) of the HCAs were published with US co-authors and their percentage in OA journals was 84%.

The similarity of the two sets of data in Figure 8 provides evidence that the two indicators, TC and U2, are closely related. After 2013, the mean value of U2 is substantially higher than TC. This suggests that this altmetric indicator could attract attention to a publication at least a year before it attracted a significant number of citations.

There are some limitations to this analysis. The biomedicine dataset is restricted to one year, and subject categories were grouped into broad areas by formal criteria. We plan to continue our analysis, investigating the influence of time-span on usage metrics. Another interesting objective of our future work is to identify hot research topics with altmetric data for Physical Chemistry.

Conclusions

We analysed a dataset of 485,505 publications, whose authors have at least one affiliation in Russia, that were indexed in the SCI-E in 2006-2019. Our objective was to explore the relationship between usage metrics (as an example of altmetrics) and classic citation metrics. Our results demonstrate a similar pattern between total citations (TC) and usage counts (U2) since 2013 in all the disciplines that we studied and in total Russian research production, except for mathematics. Chemistry, Physical, and Chemistry, Inorganic & Nuclear, have the highest values of the altmetric indicator U2 (33.4 and 21.2 respectively) in 2013 compared with any other disciplines that we studied.

The biomedicine group (6,662 publications) revealed a significant difference between the percentage of publications that did not attract any attention (15%) and the percentage of uncited publications (36%) in 2014. The share of the TC distribution has been rising because of the highly-cited articles (HCAs). It is remarkable that 28 HCAs (0.42% of the total) received 22% of all citations and 8% of the U2 counts. Only one HCA was published without international collaboration.

Since 2013 the usage count U2 values have been substantially higher than the citation score TC values. This suggests that this altmetric indicator could draw attention to a publication at least a year before it could attract a significant number of citations. Our data show that access to OA journals has an unexpected impact on usage counts U2: their values are significantly higher if the user needs to request a full-text copy from the publisher.

Altmetric indicators are now being used by many members of the Russian science community as an essential science indicator. U2 and TC do not contradict each other, but they do measure different aspects of behaviour.

Acknowledgments

We are grateful to the Russian Foundation for Basic Research (grants 20-07-00014 and 20-010-00179) for financial support, and to Clarivate Analytics for the opportunity to use their data.

References

Anon. (2008) New law mandates open access for NIH-funded research. *Analytical Chemistry*, 80(5): 1353

Begum M, Lewison G, Jassem J, *et al* (2018) Mapping cancer research across Central and Eastern Europe, the Russian Federation and Central Asia: implications for future national cancer control planning *European Journal of Cancer* 104 127–136 https://doi.org/10.1016/j. ejca.2018.08.024 PMID: 30347288

Chi, P.S. & Glänzel, W. (2019) Comparison of citation and usage indicators in research assessment in scientific disciplines and journals, *Scientometrics*, 116 (1), 537-554

Chi PS, Gorraiz J & Glänzel W. (2019) Comparing capture, usage and citation indicators: an altmetric analysis of journal papers in chemistry disciplines. *Scientometrics*, 20(3): 1461-1473

Eriksson M, Billhult A, Billhult T *et al.* (2020) A new database of the references on international clinical practice guidelines: a facility for the evaluation of clinical research. *Scientometrics*, 122:1221-1235

Grant J. (1999). Evaluating the outcomes of biomedical research on healthcare. *Research Evaluation,* 8(1): 33-38

Grant J, Cottrell R, Cluzeau F & Fawcett G. (2000) Evaluating "payback" on biomedical research from papers cited in clinical guidelines: Applied bibliometric study. *BMJ*, 320(7242(: 1107-1111

Chumachenko, A.V., Kreminskyi, B.G., Mosenkis, I.L. *et al*. (2020) Dynamics of topic formation and quantitative analysis of hot trends in physical science. *Scientometrics* 125, 739–753 https://doi.org/10.1007/s11192-020-03610-6

Fang Z., Costas R., Tian W., Wang X., Wouters P. (2020) An extensive analysis of the presence of altmetric data for Web of Science publications across subject fields and research topics. *Scientometrics* https://doi.org/10.1007/s11192-020-03564-9

Harhoff D, Narin F, Scherer FM & Vopel K. (1999) Citation frequency and the value of patneted inventions. *Review of Economics and Statistics,* 81(3): 511-515

Kramer A. (2020). Decades-Old Soviet Studies Hint at Coronavirus Strategy *The New York Times*, 24.06.20.

Kousha K, Thelwall M. .(2021) COVID-19 publications: Database coverage, citations, readers, tweets, news, Facebook walls, Reddit posts. *Quantitative Science Studies,* 1(3): 1068-1091 https://www.mitpressjournals.org/doi/full/10.1162/qss_a_00066

Lewison G., Tootell S., Roe P., Sullivan R.(2008). How do the media report cancer research? A study of the UK's BBC website . *British Journal of Cancer* 99, 569 – 576. doi:10.1038/sj.bjc.6604531

Lewison G. (2020). Web of Science usage information: a preliminary exploration of U1 and U2. Second International Conference on Science and Technology Metrics. Chinese Academy of Sciences, Zhuhai, China, 7-9 December 2020. http://socio.org.uk/stm/

Makarov V.V., Love A.J., Sinitsyna O.V. *et al* (2014) "Green" Nanotechnologies: Synthesis of Metal Nanoparticles Using Plants. *Acta Naturae,* 6 (1), 35-44 DOI: 10.32607/20758251-2014-6-1-35-44

Markusova V.A., Bogorov V.G., Libkind A.N. (2018) Usage Metrics vs Classical Metrics: Analysis of Russia's Research Productivity. *Scientometrics*, 114(2), 593-603. URL: https://doi.org/10.1007/s11192-017-2597-2

Mingels J., Leydesdorff L. (2015). A Review of Theory and Practice in Scientometrics. *European Journal of Operational Research*, 246(1): 1-19. DOI: 10.1016/j.ejor.2015.04.002 http://authors.elsevier.com/sd/article/S037722171500274X

Moed H., Markusova V., Akoev M. (2018) Trends in Russian research productivity indexed in Scopus and Web of Science. *Scientometrics*, 118 (2), 1153-1180. URL: https://doi.org/10.1007/s11192-018-2769-8.

Nalimov, V., & Mulchenko, Z. (1969), Naukometriya. Izuchenie Razvitiya Nauki kak Informatsionnogo Protsessa. [Scientometrics. Study of the Development of Science as an Information Process], Nauka, Moscow, (English translation: 1971. Washington, D.C.: Foreign Technology Division. U.S. Air Force Systems Command, Wright-Patterson AFB, Ohio. (NTIS Report No.AD735-634).

Narin F & Olivastro D. (1998) Linkage between patents and papers: An interim EPO/US comparison. *Scientometrics*, 41(1-2): 51-59

Pallari E, Eriksson M, Billhult A *et al*. (2021) Lung cancer research and its citation on clinical practice guidelines. *Lung Cancer,* 154: 44-50. DOI: 10.1016/j.lungcan.2021.01.024

Pasport nacional`nogo proekta «Nauka». Utverzhden prezidiumom Soveta pri Prezidente Rossijskoj Federacii po strategicheskomu razviti- yu i nacional`ny`m proektam (protokol ot 24 dekabrya 2018 g. N16) / Oficial`ny`j sajt Pravitel`stva Rossii. (In Russian). http://static.government.ru/media/files/vCAoi8zEXRVSuy2Yk7D8hv QbpbUSwO8y.pdf

Phillips DP, Kanter EJ, Bednarczyk B & Tastad PK (1991). Importance of the lay press in the transmission of medical knowledge to the scientific community. *New England Journal of Medicine*, 325: 1180-1183

Priem, J., Groth, P., & Taraborelli, D. (2012). The Altmetrics Collection. *PLoS ONE*, 7(11), e48753. https://doi. org/10.1371/journal.pone.0048753.

Starodubov VI, Kurakov FA, Kurakova NG, Tsvetkova LA, Polyakova YuV. (2019) Evaluating justification of choice for priority directions in the field of biomedicine in the national project «SCIENCE». *Khirurgiya. Zhurnal im. N.I. Pirogova,* 6 (1), 119-124. (In Russian). https://doi.corg/10.17116/hirurgia2019061119

Surmenev R.A., Surmeneva M.A., Ivanova A.A. (2014) Significance of calcium phosphate coatings for the enhancement of new bone osteogenesis - a review. *Acta Biomaterialia*, 10(2): 557-579

Thelwall, M., Wilson, P. (2016) Mendeley Readership Altmetrics for Medical Articles: An Analysis of 45 Fields. *Journal of the Association for Information Science and Technology*, 67(8), 1962-1972 DOI: 10.1002/asi.23501.

Widener A. (2018) European research funders demand open access. *Chemical & Engineering News*, 96(36): 14

Large coverage fluctuations in Google Scholar: a case study

Alberto Martín-Martín[1] and Emilio Delgado López-Cózar[2]

[1] *albertomartin@ugr.es*
Facultad de Comunicación y Documentación, Universidad de Granada (Spain)

[2] *edelgado@ugr.es*
Facultad de Comunicación y Documentación, Universidad de Granada (Spain)

Abstract

Unlike other academic bibliographic databases, Google Scholar intentionally operates in a way that does not maintain coverage stability: documents that stop being available to Google Scholar's crawlers are removed from the system. This can also affect Google Scholar's citation graph (citation counts can decrease). Furthermore, because Google Scholar is not transparent about its coverage, the only way to directly observe coverage loss is through regular monitorization of Google Scholar data. Thus, few studies have empirically documented this phenomenon. This study analyses a large decrease in coverage of documents in the field of Astronomy and Astrophysics that took place in 2019 and its subsequent recovery, using longitudinal data from previous analyses and a new dataset extracted in 2020. Documents from most of the larger publishers in the field disappeared from Google Scholar despite continuing to be available on the Web, which suggests an error on Google Scholar's side. Disappeared documents did not reappear until the following index-wide update, many months after the problem was discovered. The slowness with which Google Scholar is currently able to resolve indexing errors is a clear limitation of the platform both for literature search and bibliometric use cases.

Introduction

Academic bibliographic databases, and especially those that generate citation graphs, usually implement document inclusion policies that rarely allow records of documents to be removed once they have entered the system. In a bibliographic database that is intended for literature discovery, coverage stability is a desirable property, if we assume that users intuitively expect a system to retrieve the same documents over time given the same query (in addition to new documents that also meet the search criteria). This property is especially critical for some literature search use cases such as those carried out for systematic reviews, where reproducibility of the process is essential (Gusenbauer & Haddaway, 2020; Haddaway & Gusenbauer, 2020). In a citation index, the disappearance of a document would affect the citation counts of all its cited documents, impeding some types of citation analysis.

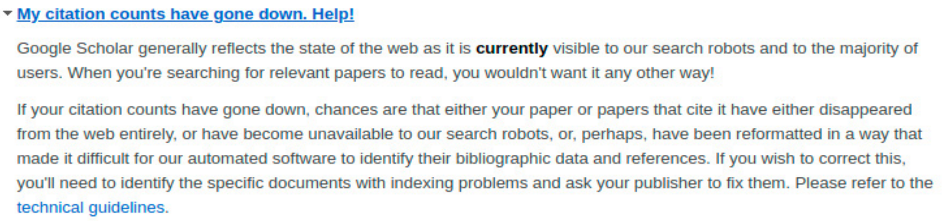
Google Scholar continues to be widely used for literature discovery, and sometimes as a data source for research evaluations. Some reasons for this are its comprehensive coverage, and that it is free to access. However, unlike other analogous tools, Google Scholar intentionally operates in a way that does not maintain coverage stability (Delgado López-Cózar et al., 2019). Instead, Google Scholar mirrors the general-purpose Google search engine, and subordinates the continued inclusion of documents in its index to their ongoing availability on the Web (as well as their continued abidance to Google Scholar's technical guidelines). Its documentation declares that this approach was chosen to provide a current reflection of the academic web at any given time (Google Scholar, n.d.-b).

As a result of this policy, coverage in Google Scholar not only increases when new documents indexed, but can also decrease when indexed documents become unavailable to Google Scholar's crawlers. To learn whether documents have stopped being available on the Web, Google Scholar carries out two complete recrawls of its index approximately every year (Google Scholar, n.d.-a). If Google Scholar's crawlers are not able to access a document during one of these recrawls, it is removed from the index.

In some cases, documents that have been removed are still findable in Google Scholar, because they are available from other sources on the Web which Google Scholar also indexes, or because they are cited in other works indexed in the system (in these cases, the document is kept as a *[CITATION]-type* record). However, in other cases, and usually after one of these major index recrawls, documents stop being findable in Google Scholar altogether, with the added consequence that citation counts of the documents that they cite also suffer a decrease (provided that cited reference metadata was available for the removed documents).

If we consider Google Scholar merely as a gateway, a digital *non-place* (Augé, 1995) through which users navigate to the places where academic documents can be accessed, rather than as a data source in its own right that could serve as a record of the inherently cumulative academic knowledge and the interactions that occur between academics, then the decision to not display information about documents that are no longer accessible is understandable (if an airport loses a flight route, it does not make sense to for it to keep displaying information about it). Furthermore, under this point of view, decreasing citation counts may not be an overly concerning issue, as their purpose would only be to serve as one of the parameters that is used to rank documents in a search, a purpose for which a certain amount of inaccuracies in citation counts can probably be tolerated.

However, Google Scholar is perceived as a bibliographic data source by many users, which makes decreasing coverage problematic. For example, researchers sometimes report author-level indicators calculated by Google Scholar in research evaluation processes (hiring, promotion, grant applications…), and seeing these figures decrease overnight and without explanation can be a cause of concern and confusion to them. Probably it is the number of questions about this phenomenon that has led Google Scholar to include a clarification in its help pages about why this occurs (Figure 1).



▾ My citation counts have gone down. Help!

Google Scholar generally reflects the state of the web as it is **currently** visible to our search robots and to the majority of users. When you're searching for relevant papers to read, you wouldn't want it any other way!

If your citation counts have gone down, chances are that either your paper or papers that cite it have either disappeared from the web entirely, or have become unavailable to our search robots, or, perhaps, have been reformatted in a way that made it difficult for our automated software to identify their bibliographic data and references. If you wish to correct this, you'll need to identify the specific documents with indexing problems and ask your publisher to fix them. Please refer to the technical guidelines.

Figure 1: Extract from Google Scholar help page that explains why citation counts of documents sometimes decrease in this platform

The issue of decreasing coverage in Google Scholar is compounded by the fact that there is no public information on the sources that are covered by this search engine. Therefore, users have no way of knowing when certain collections of documents are removed from the index, opening the possibility to situations in which users may decide to rely on this platform based on assumptions regarding its coverage that no longer hold true.

This lack of transparency means that the only way to directly observe coverage loss is through regularly monitoring Google Scholar data: recording states of the index at regular intervals. This makes this phenomenon difficult to analyse, as extracting data from Google Scholar is very time-consuming (Else, 2018), and it is difficult to anticipate which documents are going to be dropped by Google Scholar and when. Because of this, few studies have empirically documented this issue.

In November 2017, while carrying out an annual update of a directory of Spanish academic journals covered by Google Scholar Metrics (an annual product released by Google Scholar that calculates a 5-year h-index for journals), Delgado López-Cózar & Martín-Martín (2018) noticed that the number of available Spanish journals in this product had sharply decreased compared to previous years, breaking the general growing trend observed since 2012: while the 2016 edition of the directory contained 1,101 Spanish journals, in the 2017 edition only 599 journals could be found (Delgado López-Cózar & Martín-Martín, 2019). Journals from all fields disappeared, but Law journals were particularly affected, going from 156 journals in 2016 to 35 journals in 2017. Because Spanish law journals do not yet have a strong presence on the Web, we turned our attention to the largest bibliographic database that focuses on academic content published in Spain: Dialnet. This database is still the only window through which many journals published in Spain are visible on the Web, and therefore was considered likely to be involved in this *blackout* of Spanish scientific production in Google Scholar. After searching content available from Dialnet in Google Scholar and comparing the results to previous web domain analyses that we had carried in the past (Delgado López-Cózar et al., 2019), it was confirmed that most of the content from Dialnet had disappeared at some point from the search engine. This analysis was published in the Spanish LIS-focused mailing list *Iwetel,* where Dialnet's Technical Director confirmed that they had been aware of this issue since June 2017. Apparently, Google Scholar had detected that the metadata of a small batch of old records from Dialnet were inconsistent with metadata for the same documents found in other web sources, and decided to remove most of Dialnet from its index under suspicion of providing incorrect metadata. Dialnet promptly fixed this issue, but its records did not return to Google Scholar results until the following complete recrawl of the index was made public in January 2018. Thus, users interested in this content and who knew it to be covered in the past were unknowingly underserviced by Google Scholar for more than half a year. In this case, the issue was more difficult to detect because Dialnet did not contain cited reference metadata that Google Scholar could access and use in its citation graph, and therefore citation counts were not affected. This episode revealed that despite its known errors and limitations (Orduna-Malea et al., 2017), Google Scholar subjects its data to certain quality-control measures. This is, to our knowledge, the first empirically documented case of a large coverage fluctuation in Google Scholar.

In 2019, the editor of the journal *Astronomy & Astrophysics* denounced an apparently similar case (Forveille, 2019). In March 2019 the journal was notified by several researchers that citation counts to documents of this journal had decreased "by an order of magnitude" in their personal Google Scholar profiles. The editor contacted Google Scholar, who acknowledged the error and promised to remedy it. However, this would not be visible in the platform until the next complete recrawl of the index. This resulted in a sharp decrease of the h-index of this journal in the 2019 edition of Google Scholar Metrics, which was computed after Google Scholar was made aware of the issue, but using the yet uncorrected index: while the 2018 edition displayed an h5-index of 115 for this journal, in the 2019 edition this figure dropped to 52. Given the inherent resistance of a high h-index to small random changes in the underlying citation counts (probably the reason why Google Scholar favors this indicator), this signalled a very significant drop in coverage of documents in this field.

The large drop in Astronomy & Astrophysics documents in Google Scholar was also noticed by Martín-Martín et al. (2018, 2021). Analysing a collection of citations to a sample of highly-cited documents from all subject areas collected in 2018 to compare relative differences in coverage in Google Scholar, Scopus, and Web of Science, they found that Google Scholar was able to find 98% of the citations found by Web of Science, and 97% of the citations found by Scopus. Additionally, 30% of all citations were only found in Google Scholar. In 2019 the data was collected again using the same sample of highly-cited documents, but the citations to Astronomy & Astrophysics documents obtained from Google Scholar had radically changed (unlike in other subject categories, where relative differences among data sources remained mostly the same as in 2018): Google Scholar was only able to find 60% of all Web of Science citations, and 60% of all Scopus citations. Since the citation data extracted from Web of Science and Scopus in 2019 contained the same citations that were extracted in 2018, plus the new citations included in these systems between the two points of extraction, this large difference also signalled a significant drop in coverage in Google Scholar.

The datasets extracted from Google Scholar for Martín-Martín et al. (2018, 2021) provide us with an opportunity to analyse this case in more detail. Therefore, the goal of this study is to document this case, to try to find out the cause of this sudden drop in coverage to documents in the field of Astronomy & Astrophysics, and to check whether this issue was resolved in subsequent recrawls of Google Scholar's index.

Methods

The datasets extracted from Google Scholar and analyzed in Martín-Martín et al. (2018, 2021) were used. These datasets contain the lists of citing documents to a sample of 2515 highly-cited documents from 2006 that Google Scholar released in 2017 with the name *Google Scholar Classic Papers* (GSCP). In this product[1], Google Scholar displayed the top 10 most cited documents published in 2006 in each of 252 subject categories. For more information about this product, see Orduna-Malea et al. (2018).

Since in this study we are particularly interested in the coverage of Astronomy & Astrophysics documents in Google Scholar, only the 10 highly-cited documents in this field in GSCP (Table 1), and the list of citing documents in Google Scholar for each of these documents, are analyzed here.

The list of documents that cite each document in Table 1 was extracted from Google Scholar in three different occasions: April-May of 2018, May-June 2019, and April 2020. To do this, a custom scraper was used (Martín-Martín, 2018).

The metadata extracted from Google Scholar for each of these citing documents was enriched by complementing it with metadata available in the HTML meta tags of the webpages where Google Scholar found these documents, and the metadata available in CrossRef's and arXiv's public APIs. A DOI was found for 79% of the citations in the 2018 dataset, 76% of the citations in the 2019 dataset, and 77% of the citations in the 2020 dataset.

[1] https://scholar.google.com/citations?view_op=list_classic_articles&hl=en&by=2006

Table 1: Highly-cited documents in Astronomy & Astrophysics in Google Scholar Classic Papers

Art. #	*Journal*	*Vol(issue), pages*	*DOI*
1	The Astronomical Journal	131(2), 1163-1183	10.1086/498708
2	Monthly Notices of the Royal Astronomical Society	365(1) 11-28	10.1111/j.1365-2966.2005.09675.x
3	Astronomy & Astrophysics	447(1), 31-48	10.1051/0004-6361:20054185
4	Monthly Notices of the Royal Astronomical Society	370(2) 645-655	10.1111/j.1365-2966.2006.10519.x
5	Nuclear Physics A	777, 1-4	10.1016/j.nuclphysa.2005.06.010
6	The Astrophysical Journal	651(1), 142-154	10.1086/506610
7	Physical Review D	74(12), 123507	10.1103/PhysRevD.74.123507
8	The Astronomical Journal	131(4), 2332-2359	10.1086/500975
9	Monthly Notices of the Royal Astronomical Society	368(1) 2-20	10.1111/j.1365-2966.2006.10145.x
10	Monthly Notices of the Royal Astronomical Society	372(3) 961-976	10.1111/j.1365-2966.2006.10859.x

Citations across the three datasets were matched based on (in this order) Google Scholar's internal document identification codes, the URLs of the webpages were Google Scholar found the citing documents, the DOIs of the citing documents, and a combination of title and author similarity, in a similar way as described in Martín-Martín et al. (2018, 2021):

1. For each pair of datasets, A and B and a seed highly-cited document X, all citing documents with a document id (Google Scholar ID, URL, and DOI) that cite X according to A were matched to all citing documents with a corresponding document id that cite X according to B.
2. For each of the unmatched documents citing X in A and B, a further comparison was carried out. The title of each unmatched document citing X in A was compared to the titles of all the unmatched documents citing X in B, using the restricted Damerau-Levenshtein distance (optimal string alignment) (Damerau, 1964; Levenshtein, 1966). The pair of citing documents which returned the highest title similarity (1 is perfect similarity) was selected as a potential match. This match was considered successful if either of the following conservative heuristics was met:
    1. The title similarity was at least 0.8, and the title of the citing document was at least 30 characters long (to avoid matches between short, undescriptive titles such as "Introduction").
    2. The title similarity was at least 0.7, and the first author of the citing document was the same in A and B.

First, citations in the 2019 dataset were matched to citations in the 2018 dataset. The result of that matching was in turn matched to the citations in the 2020 dataset.

Results

A simple observation of the citation counts reported by Google Scholar over the years for the 10 highly-cited documents in the field of Astronomy & Astrophysics in GSCP already reveals a large fluctuation (Figure 2).

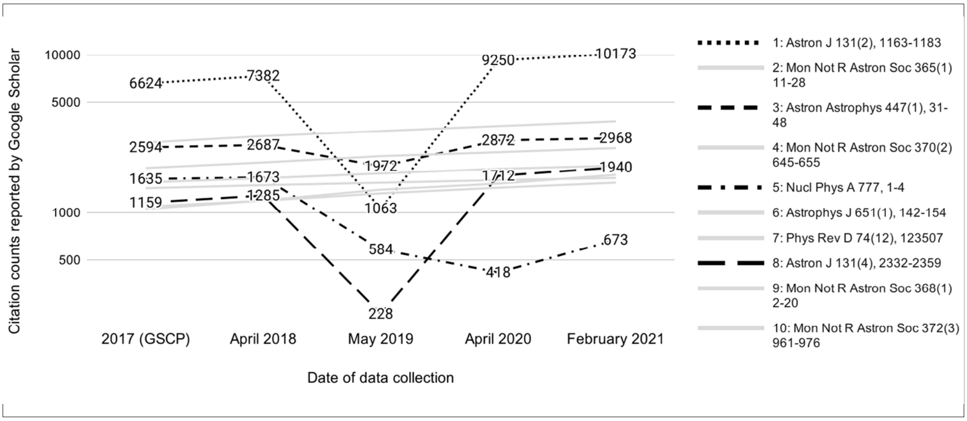


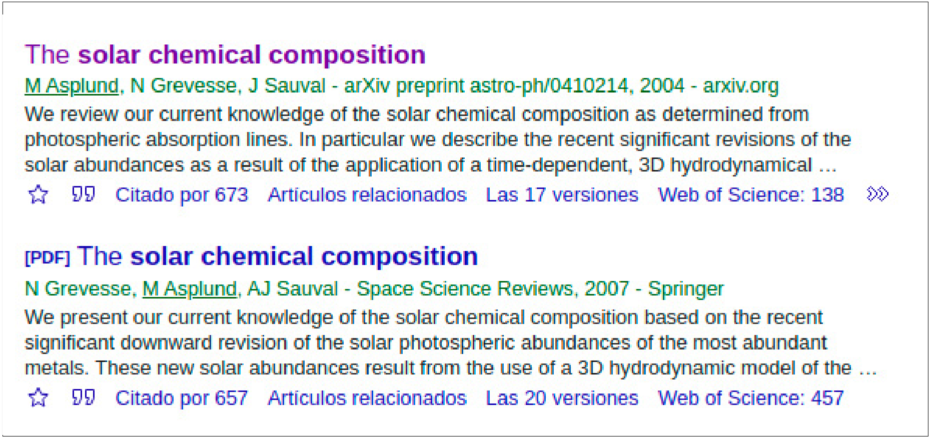
Figure 2: Citation counts reported by Google Scholar over the years to the 10 highly-cited documents in the field of Astronomy & Astrophysics in GSCP.

Four of the 10 documents suffered a sharp decrease in citations in 2019. In the most extreme case (document #1), the citation count in 2019 had decreased by 6,319 citations respect to the count reported in 2018. It is important to note that this document was published not in the journal *Astronomy & Astrophysics*, but in *The Astronomical Journal*, published by IOP Science. Indeed, of the four documents that show a drop in citation counts in 2019, only one was published in *Astronomy & Astrophysics* (document #3). There is a second article from *The Astronomical Journal* with decreased citation counts in 2019 (document #8), and another one published *Nuclear Physics A* (document #5). This shows that *Astronomy & Astrophysics* was not the only journal to be affected by the drop in coverage in 2019, which is expected, since whichever documents disappeared from Google Scholar likely cited articles from various journals in the field.

Six of the ten documents (#2, #4, #6, #7, #9, #10), however, do not show clear signs of a coverage drop: their citation counts grew each year. Four of these documents were published in *Monthly Notices of the Royal Astronomical Society* (Oxford University Press), one in The Astrophysical Journal (IOP Science), and one in Physical Review D (American Physical Society). This does not rule out the possibility that they lost citations in 2019, only that the growth/loss balance was positive. Nevertheless, this suggests that some documents could have been more affected than others.

Of the four documents that clearly lost citations in 2019, three of them seem to recover them by 2020 and continue receiving citations in 2021. This suggests that the coverage loss was indeed temporary, and that citations were recovered at some point between summer of 2019 and spring of 2020, probably after the second complete recrawl of the index in 2019. One of the documents (#5), however, does not recover the amount of citation counts that were reported in 2017 and 2018, even by 2021. After closer examination of this document, titled "The solar chemical composition" and published in *Nuclear Physics A*, it was discovered that there are actually two documents with the same title and the same authors, published in two different journals (Figure 3). We believe it is probable that in 2017 and 2018, these two documents were incorrectly merged into one record (combining the citations of the two), and that in 2019 they were separated, as they remain in 2021. Since this is a different kind of Google Scholar error than those we are documenting here, this document and its citations were excluded from further analysis in this study.

We analysed the list of documents that cited each of the nine highly-cited document according to Google Scholar at three points in time: 2018, 2019, and 2020. In 2018, 21,907 citations were extracted. In 2019, 15,042 were extracted, while in 2020, 25,195 citations were found in Google Scholars to these nine documents. Of the 21,907 citations found in 2018, 8,840 (40%) were missing in 2019. In 2020, however, 96% of the citations available in the 2018 dataset had reappeared.



The **solar chemical composition**

M Asplund, N Grevesse, J Sauval - arXiv preprint astro-ph/0410214, 2004 - arxiv.org

We review our current knowledge of the solar chemical composition as determined from photospheric absorption lines. In particular we describe the recent significant revisions of the solar abundances as a result of the application of a time-dependent, 3D hydrodynamical ...

☆ 99 Citado por 673 Artículos relacionados Las 17 versiones Web of Science: 138 ≫

[PDF] The **solar chemical composition**

N Grevesse, M Asplund, AJ Sauval - Space Science Reviews, 2007 - Springer

We present our current knowledge of the solar chemical composition based on the recent significant downward revision of the solar photospheric abundances of the most abundant metals. These new solar abundances result from the use of a 3D hydrodynamic model of the ...

☆ 99 Citado por 657 Artículos relacionados Las 20 versiones Web of Science: 457

Figure 3: Two documents with the same title and authors, published in different journals. It is possible that Google Scholar incorrectly merged these two records into one (combining citation counts) in 2017 and 2018, and separated them in 2019. Screenshot taken on 14/02/2021.

To find out exactly which documents caused the decrease in citation counts, the citations found in 2018 to each of the nine highly-cited documents were grouped by the publisher of the citing document, and by whether or not the citation was also found in 2019 and 2020 (Figure 4). Missing documents are mostly concentrated in document #1, #8, and #3, while the other six documents are affected to a much lower extent.

Missing documents in 2019 are distributed across many of the largest publishers in the field (Figure 4). They are also distributed across all publication years. Therefore, because we know that these large publishers did not actually disappear from the Web in 2019, and it is unlikely that they concertedly modified the metadata of their documents or their crawler access rules so as to not comply with Google Scholar's technical guidelines, it is likely that the error originated at Google Scholar's side. The exact cause of the error is, nevertheless, unknown.

There are several possibilities as to why some citations from 2018 are still missing in the 2020 dataset: some may correspond to documents that did actually disappear from the web (particularly those in the category "Other", in which some of the documents are hosted in less stable websites). Some others may correspond to duplicate records found in 2018 that were subsequently merged.

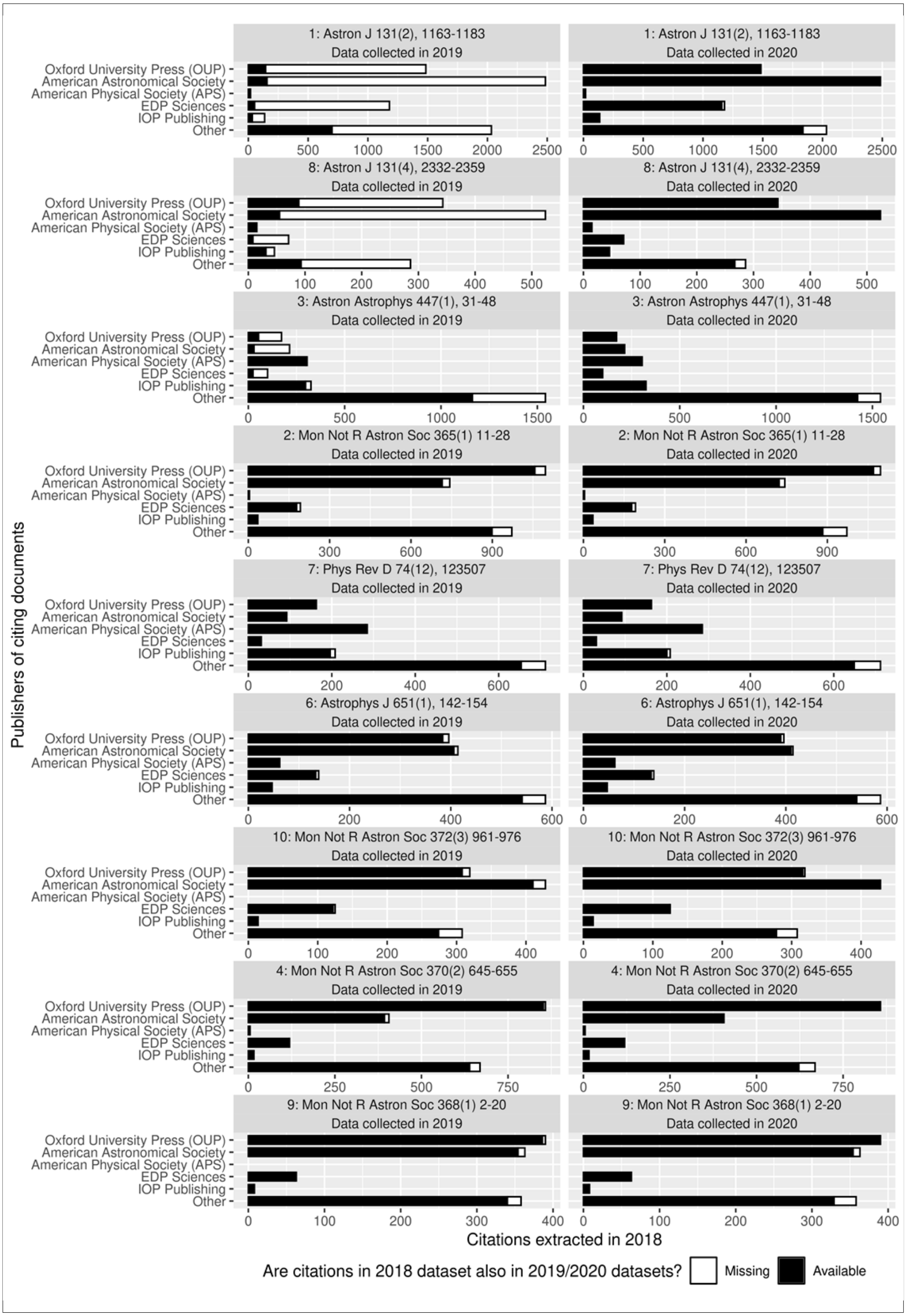


Figure 4: Citations found in 2018 for each of the 9 highly-cited documents, grouped by publisher of citing document and by whether or not the citation was found in 2019 and 2020.

The citing documents that were present in the three datasets (data extracted in 2018, 2019, and 2020) came with citation counts of their own, which provides us with an opportunity to gauge to what extent each publisher was affected by the loss in coverage in 2019 using a larger sample than the 10 highly-cited documents in GSCP. In this regard, documents published by EDP Sciences (publisher of the journal *Astronomy & Astrophysics*) were severely affected by the loss of coverage (Figure 5), followed by documents published by the American Astronomical Society, which were affected to a much lower extent. Out of the 724 documents published by EDP Sciences that were available in the three datasets, 404 of them (58%) reported at least 10 citations less in the 2019 dataset than in the 2018 dataset, whereas out of the 2,604 documents published by the American Astronomical Society present in the three datasets, 141 (5%) reported at least 10 citation less in the 2019 dataset than in the 2018 dataset. In the rest of publishers, lower citations in the 2019 dataset than in the 2018 dataset were even more uncommon.

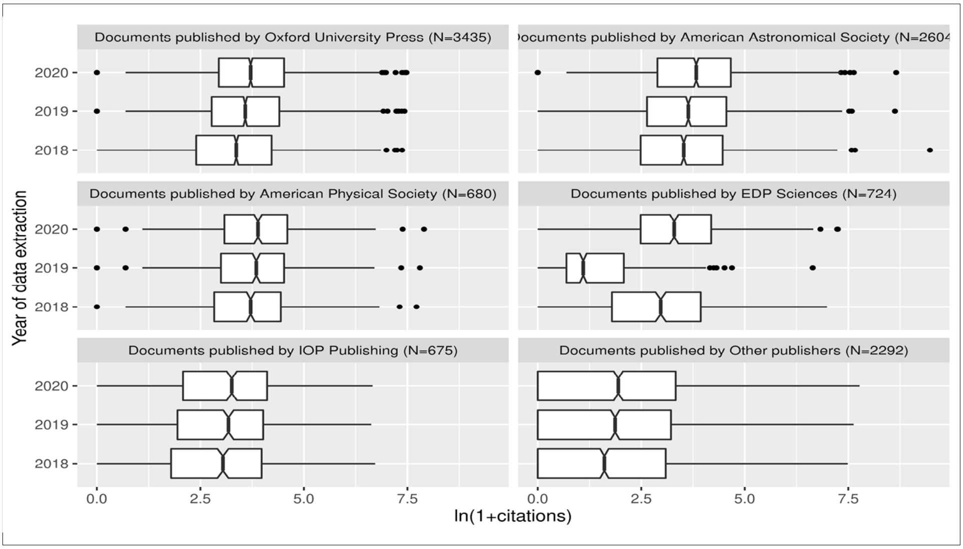


Figure 5: Distribution of (log-transformed) citation counts of citing documents in 2018, 2019, and 2020, grouped by publisher.

Discussion and conclusions

The results confirm that this is another case of a large coverage fluctuation in Google Scholar. The only other documented case of a coverage fluctuation at a similar scale is the one that affected Dialnet in 2017 (Delgado López-Cózar & Martín-Martín, 2018). Some similarities and differences between the two cases are drawn here:

- In the case described here, the effects were quickly reported by researchers who noticed drastic drops in citation counts in documents of the area, especially in the journal *Astronomy & Astrophysics*. On the other hand, in the Dialnet event, the effects of the coverage drop did not draw much attention until a few months after the coverage drop, when many Spanish journals were found to have disappeared from Google Scholar Metrics (although Dialnet's managers declared to be aware of the issue from the beginning). Dialnet records do not provide citation data, which could explain why the effects on the surrounding network were less noticed.

- In the case described here, we have only ascertained that a large number of citation connections were dropped by Google Scholar, causing sudden and sharp decreases of citation counts in other documents. It is not clear whether the citing documents also became unavailable in document searches, as in the Dialnet event. If this was the case, it would mean that users interested in Astronomy and Astrophysics content during 2019 who expected Google Scholar to have a comprehensive coverage of this field (based on previous experiences with the platform) could have been unknowingly underserviced for a period of 6 to 9 months.
- In the event that affected Dialnet, the managers of this database explained that the problem was caused by a small batch of old records with an incorrect author order. Google Scholar detected the difference in metadata between Dialnet and other sources that covered the same documents, and decided to quarantine metadata from Dialnet until the issue was resolved. In the case studied here, the direct cause of the issue has not been made public by either Google Scholar or the affected journals.

Although there are reports that Google Scholar sometimes contacts content providers when an issue like this arises (Delgado López-Cózar & Martín-Martín, 2018), the slowness with which it is currently able to resolve this kind of issues is a clear limitation of the platform for literature search use cases as well as for bibliometric use cases.

This case highlights the tension that is created when design choices that make sense in some scenarios, such as continued online availability as an inclusion criterion in general purpose search engines, are applied to more restricted domains, like academic literature search, where users might have different expectations (i.e. for content to be stable and to grow over time). These assumptions are further reinforced when other services in which stability is also to be expected (journal rankings, author profiles) are built on top of the original data.

Anurag Acharya, chief engineer of Google Scholar, once declared that, unlike the case of searches in the general Google, "the world of research is global", which was the reason why Google Scholar decided to implement a relevance ranking that does not take into account data from a user's previous searches (ALPSP, 2015). Events like the one documented here make the case to also consider the expectation of coverage stability and continued growth as one of the properties that characterises academic literature searches and therefore one of the principles that should guide the design choices of academic search tools.

References

ALPSP. (2015, September 21). Co-founder of Google Scholar Anurag Acharya asks What happens when all articles are easy to find? https://youtu.be/S-f9MjQjLsk?t=3342

Augé, M. (1995). *Non-places: Introduction to an Anthropology of Supermodernity*. Verso.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3), 171–176. https://doi.org/10.1145/363958.363994

Delgado López-Cózar, E. & Martín-Martín, A. (2018). Apagón digital de la producción científica española en Google Scholar. *Anuario ThinkEPI*, 12, 265–276. https://doi.org/10.3145/thinkepi.2018.40

Delgado López-Cózar, E. & Martín-Martín, A. (2019). *Indice H de las revistas cientificas españolas en Google Scholar Metrics 2014-2018*. https://doi.org/10.13140/RG.2.2.36649.13923

Delgado López-Cózar, E., Orduna-Malea, E. & Martín-Martín, A. (2019). Google Scholar as a data source for research assessment. In W. Glänzel, H. Moed, U. Schmoch & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators*. Springer.

Else, H. (2018, April 11). How I scraped data from Google Scholar. *Nature(News)* https://doi.org/10.1038/d41586-018-04190-5

Forveille, T. (2019). A&A ranking by Google. *Astronomy & Astrophysics*, *628,* E1. https://doi.org/10.1051/0004-6361/201936429

Google Scholar. (n.d.-a). *Search Tips: Content coverage*. Google Scholar. Retrieved December 2, 2021, from https://web.archive.org/web/20210212175858/https://scholar.google.com/intl/es/scholar/help.html#coverage

Google Scholar. (n.d.-b). *Search Tips: Inclusion and corrections*. Google Scholar. Retrieved December 2, 2021, from https://web.archive.org/web/20210212175858if_/https://scholar.google.com/intl/es/scholar/help.html#corrections

Gusenbauer, M. & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, 11(2), 181–217. https://doi.org/10.1002/jrsm.1378

Haddaway, N. & Gusenbauer, M. (2020, February 3). A broken system – why literature searching needs a FAIR revolution. *Impact of Social Sciences*. https://blogs.lse.ac.uk/impactofsocialsciences/2020/02/03/a-broken-system-why-literature-searching-needs-a-fair-revolution/

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707–710.

Martín-Martín, A. (2018). *Code to extract bibliographic data from Google Scholar*. https://doi.org/10.5281/zenodo.1481076

Martín-Martín, A., Orduna-Malea, E., Thelwall, M. & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

Martín-Martín, A., Thelwall, M., Orduna-Malea, E. & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1), 871–906. https://doi.org/10.1007/s11192-020-03690-4

Orduna-Malea, E., Martín-Martín, A. & Delgado López-Cózar, E. (2017). Google Scholar as a source for scholarly evaluation: A bibliographic review of database errors. *Revista Española de Documentación Científica*, 40(4), e185. https://doi.org/10.3989/redc.2017.4.1500

Orduna-Malea, E., Martín-Martín, A. & Delgado López-Cózar, E. (2018). Classic papers: Using Google Scholar to detect the highly-cited documents. *23rd International Conference on Science and Technology Indicators*, 1298–1307. https://doi.org/10.31235/osf.io/zkh7p

Brazilian scientific production on COVID-19: a bibliometric and altmetric analysis

Isabella Maria Almeida Mateus[1] and Cristian Berrío-Zapata[2]

[1] *isabellamateus@iec.gov.br*
Universidade Federal do Pará, Instituto de Ciências Sociais Aplicadas, Programa de Pós-Graduação em Ciência da Informação, Rua Augusto Corrêa, 01 - Guamá. CEP 66075-110 Belém, Pará (Brazil)
Instituto Evandro Chagas/SVS/MS, Rodovia BR-316 Km 7, s/n - Levilândia. CEP 67030-000 Ananindeua, Pará (Brazil)

[2] *berriozapata@ufpa.br*
Universidade Federal do Pará, Instituto de Ciências Sociais Aplicadas, Programa de Pós-Graduação em Ciência da Informação, Rua Augusto Corrêa, 01 - Guamá. CEP 66075-110 Belém, Pará (Brazil)

Abstract

We analyze the Brazilian production and diffusion of scientific literature on COVID-19 in 2020, to assess what was the protagonism of this country during this event. All documents were obtained from Web of Science and considered under bibliometric and altmetric analysis. We used the production and citation indicators crossed with other variables like author, institutional affiliation, impact factors, funding organizations, research areas, etc. A total of 2,703 articles were found, mostly published in December and in English, having received 10,190 citations, with an average of 3.77 citations per document, a value lower than that found for other countries. Ten journals concentrated publications, being only one of those not Brazilian. São Paulo University was the most producing institution. Analyzing the top 100 most cited articles with an altmetric tool, the average Altmetric Attention Score (AAS) was 392.22. From the 40,062 mentions on social media, 94.8% came from Twitter, mainly by general public (78.6%). The top 10 articles with the highest AAS concentrated 79,5% of all attention; eight of them were original investigations. We propose a content analysis to verify which specific topics are attracting greater interest from the general public, but not so much from the scientific community.

Introduction

The new coronavirus emergence caused a rapid and massive growth in research at a fast pace, ending in a scientific race. The challenge was to rapidly identify relevant, accurate, and robust research. COVID-19 effects were widely discussed in digital social media, so science produced a mediatic public debate in real time. The response of scientific communities to epidemics such as H1N1, Zika, Ebola, MERS, and SARS, showed increased production immediately after the outbreak declaration by WHO. For COVID-19, the official pandemic declaration generated a blast in scientific publications.

Brazil has become a major player in global science, although it is currently suffering a severe crisis affecting its scientific community. The country is part of the group of nations that leads publishing on COVID-19. Several authors have analyzed this publication with traditional bibliometrics, or during short periods of time, but these studies do not show the conversation between science and society, expressed by the interaction between bibliometric and altmetric indicators. In this work, we will characterize the COVID-19 Brazilian production using one full year of bibliometric indicators and will cross this result with altmetric indicators within social media.

As a result, this article shows the most relevant characteristics of the Brazilian COVID-19 production and dissemination to society through social media during 2020.

Context

Since the emergence of the new coronavirus much knowledge has been mapped. The rapid spread of the COVID-19 across the globe has caused the number of investigations related to the disease to grow, also at a very fast pace, setting an unprecedented scientific race (Torres-Salinas, 2020).

The challenge for scientists – especially at the beginning of an epidemic – is to identify relevant research results in a short amount of time, when the availability of accurate and robust clinical, laboratory, and epidemiological data is essential to guide decision-making in a timely manner, especially concerning public health issues (Xu et al., 2020). Therefore, demands for speed and efficiency in the dissemination of scientific results impose an additional challenge to the scientific communication system (Larivière, Shu & Sugimoto, 2020; Torres-Salinas, Robinson-Garcia & Castillo-Valdivieso, 2020).

The outbreak has highlighted the relevance of data sharing, developments, and research results in a faster (and more open) way than the traditional scientific communication system (Larivière, Shu & Sugimoto, 2020). In this race imposed by the pandemic, the response from editors has been efficient, as scientific journals are giving priority to works related to COVID-19 (Horbach, 2020), reducing the average publication time, making available articles in preprint (not yet peer-reviewed) and open access, like in other historical public health crises, such as the Zika virus outbreak in Brazil (Araujo et al., 2017).

COVID-19 effects are vividly discussed in digital environments like platforms and social media such as Twitter (Chen, Lerman & Ferrara, 2020; Melo & Figueiredo, 2020), so this is the first time that a pandemic can be described, debated, and investigated by the scientific community in real time and online, by means considered conventional (journal articles) in association with social media (Boetto et al., 2020).

Initiatives such as those undertaken by publishers and the strengthening of interaction on social networks around scientific research go toward Open Science, a comprehensive term that involves sharing practices in open access, but also calls for greater transparency and valuing the participation of nonscientists and non-specialists in the production and dissemination of knowledge, constituting a "new scientific practice" (Oliveira & Silva, 2016).

Studies analyzing the response of the scientific community to international public health emergencies, such as H1N1, Zika, Ebola, MERS, and SARS, showed that the number of scientific publications increased immediately after the outbreak was declared by WHO (Haghani et al., 2020; Zhang et al., 2020). In the case of COVID-19, it was no different. The official declaration of the pandemic generated an unprecedented growth in scientific publications from different areas of knowledge (Brainard, 2020).

Brazil, although suffering severe financial and infrastructure cuts affecting its scientific community (Norte, 2020; De Negri et al., 2020), still is part of the group of countries that leads publishing on COVID-19, since the beginning of the outbreak (Bernardes & Dorado, 2020).

Several authors have presented analyzes on COVID-19 publication using traditional bibliometric indicators (Chahrour et al., 2020; Figueredo et al., 2020; Oliveira et al., 2021; Silva et al., 2020). However, these studies do not show an association between bibliometric and altmetric indicators, from the perspective of Brazilian scientific production on COVID-19, nor do they undertake an analysis of the impact and diffusion of these publications using altmetrics.

Within this context, this study analyzes Brazilian COVID-19 production and dissemination of knowledge, identifying the most relevant characteristics of this domain, to provide an overview of the impact of this literature in social media during 2020.

Methods

This is a case study about Brazil's scientific publication behavior during the first year of the COVID-19 outbreak. We used both bibliometric and altmetric indicators, to identify the diffusion and impact of Brazilian academic publications related to the new coronavirus. Traditional bibliometric indicators can be complemented with altmetrics. Bibliometrics allows analyzing the dynamics of publications during a certain period of time, while altmetrics let us apprehend people's interaction on the social web around research products and assess the wider impact of science following online activities and tools (Baheti & Bhargava, 2017; Piwowar, 2013; Priem et al., 2010; Williams, 2017). The combination of methods allows updating the concept of scientific influence (Barros, 2015). Alternative metrics are important for the analysis of emerging online conversations of global importance, such as COVID-19, since they have indicators of scientific production, dissemination, and appropriation beyond the academic environment and its typical channels of dissemination (Borrego, 2014; Williams, 2017).

Scientific articles, published between January 2020 and December 2020, were collected from the Web of Science (WoS) Core Collection database (Clarivate Analytics, Philadelphia, PA, USA) on January 25, 2021. Descriptors were "COVID-19" and "SARS-CoV-2". WoS includes some of the finest and more robust databases about the health area; its results are comparable to SCOPUS database, but in the case of Brazil we found 20% additional documents. All articles were analyzed using descriptive bibliometrics to have a profile of their main characteristics in production and citations. The results were organized following selected variables: author, institutional affiliation, journal, journal impact factor (JIF), language, document type, publication data, citation level, access, funding organizations, and research areas.

The top 100 articles with the highest citation were analyzed using the Bookmarklet for Researchers (Altmetric, London, UK) altmetric tool, which provides, among other data, the Altmetric Attention Score (AAS). We assumed that, following Lotka’s production principle and Merton’s Matthew Effect, that less than 5% of all publishing authors and papers, would accumulate the majority (more than 70%) of citations or “visibility”. AAS weights the attention driven by a scientific article in social media, reporting a compound indicator reflecting the number of mentions and shares in news, documents about public policy, blogs, contents at Twitter, Facebook, Wikipedia, Reddit, references in videos, and Mendeley readers, all packed into an index.

Results

The search in WoS returned 2,703 records that received a total of 10,190 citations, an average of 3.77 citations per document. It was possible to observe that most papers were published in December 2020 (312; 17.5%) and September 2020 (275; 15.5%), as seen in Figure 1.

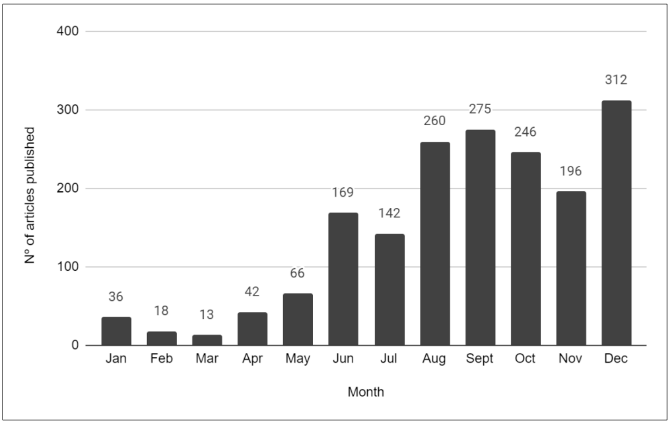


Figure 1. Number of articles authored/co-authored by Brazilian researchers on COVID-19 published per month, during the year 2020 (Source: WoS, 2021).

Most frequent documents were original articles 1,458 (53.9%), followed by letters 413 (15.3%), editorials 391 (14.5%), and review articles 390 (14.4%). Most documents were in English (2,381; 88.1%), followed by Portuguese (294; 10.9%). The number of institutions/authorships reported in WoS as involved in the production of these documents was 5,010. University of São Paulo (USP) stood as the most productive institution with 583 articles (21.6%), followed by the Federal University of São Paulo (UNIFESP) with 222 (8.2%), and Oswaldo Cruz Foundation with 216 (8.0%). Table 1 highlights the 10 most producing institutions.

Table 1. The most productive institutions of Brazilian authors/co-authors who published articles on COVID-19 in 2020 (Source: WoS, 2021).

Institution	*Country*	*Articles*	
		n	*% of 2,703*
University of São Paulo	Brazil	583	21.6
Federal University of São Paulo	Brazil	222	8.2
Oswaldo Cruz Foundation	Brazil	216	8.0
Federal University of Rio de Janeiro	Brazil	179	6.6
University of Campinas	Brazil	142	5.2
Federal University of Minas Gerais	Brazil	138	5.1
Federal University of Rio Grande do Sul	Brazil	105	3.9
University of London	England	90	3.3
Federal University of Bahia	Brazil	85	3.1
Federal University of Pernambuco	Brazil	85	3.1

Institutions/funding agencies for research were mentioned 2,105 times in 623 articles. From these, 1,339 mentions were about Brazilian institutions. The Brazilian National Council of Scientific and Technological Development (CNPq) was the most prevalent, with 400 (19.0%) mentions, followed by the Coordination of Superior Level Staff Improvement (CAPES), with 318 (15.1%), and São Paulo State Research Support Foundation (FAPESP), with 174 (8.7%) occurrences. About 77% (2,080) of the records did not contain any data regarding research funding.

A total of 15,359 researchers were counted, among Brazilian and foreign authors and co-authors. The most productive author/co-author published 19 articles, while the majority, 12,668 (82.5%) published only one article. The 10 most productive Brazilian researchers are shown in Table 2.

Table 2. The most productive Brazilian authors/co-authors who published articles on COVID-19 in 2020 (Source: WoS, 2021).

Author	*Institution*	*Articles*	
		n	*% of 2,703*
Carvalho, WB	University of São Paulo, SP	19	0.70
Souza, CDF	Federal University of Alagoas, AL	17	0.63
Rocco, PRM	Federal University of Rio de Janeiro, RJ	16	0.59
Santos, VS	Federal University of Alagoas, AL	15	0.56
Kowalski, LP	University of São Paulo, SP	14	0.52
Martelli Jr., H	State University of Montes Claros, MG	14	0.52
Martins-Filho, PR	Federal University of Sergipe, SE	14	0.52
Marchiori, E	Federal University of Rio de Janeiro, RJ	13	0.48
Rolim Neto, ML	Federal University of Cariri, CE	13	0.48
Giovanetti, M	Oswaldo Cruz Foundation, RJ	12	0.44

The geographic distribution of retrieved publications shows the collaboration of Brazil with 134 countries. The spread of international collaboration includes the US in 461 documents

(17.0%), England in 244 (9.0%), Italy in 218 (8.1%), Spain in 183 (6.8%), and Canada in 151 (5.6%).

Table 3. Characteristics and components of the top 100 most cited COVID-19 articles published by Brazilian authors with the highest Altmetric Attention Scores, 2020.

Characteristic	*Value*
Altmetric Attention Score, median (range)	392.22 (1–6889)
Traditional citation count, median (range)	58.2 (16–551)
Journal impact factor, median (range)	11.487 (0–74.699)
News mentions, total (range)	1582 (0–410)
Blog mentions, total (range)	222 (0–35)
Policy mentions, total (range)	36 (0–6)
Twitter mentions, total (range)	37965 (1–7536)
Demographic breakdown, total (range)	
Members of the public	29830 (0–6291)
Scientists	3944 (0–653)
Practitioners	3289 (0–599)
Science communicators	889 (0–195)
Unknown	13 (0–2)
Geographical breakdown, n (%)	
USA	36 (36)
United Kingdom	15 (15)
Brazil	12 (12)
Spain	6 (6)
Mexico	4 (4)
Japan	3 (3)
Others	24 (24)
Facebook mentions, total (range)	177 (0–24)
Wikipedia mentions, total (range)	18 (0–6)
Reddit mentions, total (range)	43 (0–7)
Video mentions, total (range)	19 (0–5)
Mendeley readers, total (range)	39980 (77–2234)
Researcher, total (range)	5023 (9–273)
Bachelor student, total (range)	5212 (0–379)
Postgraduate student, total (range)	443 (0–62)
Master student, total (range)	4575 (0–328)
Doctoral student, total (range)	1010 (0–137)
PhD student, total (range)	2651 (0–324)
Professor/Lecturer, total (range)	82 (0–24)
Other (e.g., Librarian), total (range)	19227 (0–707)
Article type, n (%)	
Original investigation	49 (49)
Editorial	20 (20)
Letter	13 (13)
Review	18 (18)
Journal’s knowledge area (n=79), n (%)	
Biomedicine	62 (78.5)
Psychiatry & Behavioral Sciences	8 (10.1)
Public Health	4 (5.0)
Physics & Mathematics	2 (2.5)
Environmental Sciences	1 (1.3)
Interdisciplinarity	1 (1.3)
Virology	1 (1.3)

There were 122 research areas identified. Half of them (67; 54.9%) were related to human health; recurrent topics were “Public Environmental Occupational Health” (335; 12.4%), “General Internal Medicine” (309; 11.4%), “Infectious Diseases” (139; 5.1%), “Neurosciences Neurology” (112; 4.1%), and “Science Technology Other Topics” (112; 4.1%).

The articles were distributed in 991 scientific journals, with an average of 2.73 articles per journal; 898 (96.6%) were international and 93 (9.4%) Brazilian. Of the 10 journals with the highest number of publications, only one was not Brazilian: CLINICS, 74 (2.7%); Cadernos de Saúde Pública, 68 (2.5%); Revista da Associação Médica Brasileira, 68 (2.5%); Ciência & Saúde Coletiva, 58 (2.1%); Revista da Sociedade Brasileira de Medicina Tropical, 47 (1.7%); Arquivos Brasileiros de Cardiologia, 46 (1.7%); Revista de Administração Pública, 43 (1.6%); Revista Tecnologia e Sociedade, 34 (1.2%); PLOS ONE, 28 (1.0%); and Vigilância Sanitária em Debate: Sociedade, Ciência & Tecnologia, 27 (1.0%). The number of journals that published a single document was 613 (61.9%).

As for the altmetric analysis of the 100 most cited articles, most of them were published in biomedical journals (62/79; 78.5%); however, compared to other journals, the Science of the Total Environment published the largest number of articles (6/100; 6.0%). All articles were open access; only 49.0% were original investigations, and 43.0% of the studies were considered Brazilian because they presented the majority or the totality of authors from Brazilian institutions. To analyze the real contribution of original investigations from Brazil, we cross-referenced how many studies were of this type against how many were Brazilian – by the criteria above –, resulting in 57.1%. The medians of AAS, citations per document, and JIF were 392.22, 58.5, and 11.487, respectively (Table 3).

A total of 40,062 mentions on social media platforms were found. The highest number of mentions for the selected articles was found on Twitter (37,965/40,062; 94.8%). The geographic analysis of the tweets showed that US (36.0%), United Kingdom (15.0%), and Brazil (12.0%) had the highest number of mentions. The demographic analysis of all tweets showed that 78.6% were from members of the general public, 10.4% from scientists, 8.7% from professionals (doctors, other health professionals), and 2.3% from scientific communicators (journalists, bloggers, editors). Altogether, there were 39,980 mentions of Mendeley, most of which were identified as coming from undergraduate students (5,212; 13.0%), researchers (5,023; 12.6%), and master's students (4,575; 11.4%) (Table 3).

Table 4. Top 10 COVID-19 articles published by Brazilian authors in 2020 by AAS and its position in the ranking of the 100 most cited articles according to WoS, Jan./2021.

Article	*Source*	*JIF*	*AAS*		*Citations*	
			n1	*#*	*n2*	*#*
Cavalcanti et al., 2020	N Engl J Med	74.699	6.889	1	110	12
Sterne et al., 2020	JAMA	45.540	5.961	2	79	20
Emanuel et al., 2020	N Engl J Med	74.699	5.043	3	551	1
Siemieniuk et al., 2020	BMJ	30.313	2.942	4	23	45
Borba et al., 2020	JAMA Netw Open	5.032	2.594	5	327	3
Van Bavel et al., 2020	Nat Hum Behav	12.282	1.968	6	335	2
Bastos et al., 2020	BMJ	30.313	1.911	7	72	22
Tomazini et al., 2020	JAMA	45.540	1.810	8	40	32
Candido et al., 2020	Science	41.846	1.116	9	21	47
Codo et al., 2020	Cell Metab	21.567	962	10	21	47

JIF = Journal Impact Factor; *AAS* = Altmetric Attention Score; *n1* = score value; *n2* = times cited; # = ranking position.

The top 10 articles with the highest AAS concentrated 79,5% of all registered attention; eight of them were original investigations; and four of them were considered Brazilian studies by the criteria described above. The article with the highest AAS (6,889) addressed the use of

Hydroxychloroquine with or without Azithromycin in patients with mild to moderate COVID-19 (Cavalcanti et al., 2020) and was published in the New England Journal of Medicine (JIF 74.699). Despite occupying the 1st position in the ranking of articles that received the highest index of attention in social media, this article was only in 12th position among the most cited. The article by Emanuel et al. (2020) was the most cited and performed well among those who had higher AAS, occupying the 3rd position among articles published by Brazilian authors/co-authors. None of the journals with the highest AAS were Brazilian and 70.0% had JIF ≥ 30.0 (Table 4).

Discussion

Through this study, it was possible to verify the characteristics of Brazilian scientific production on COVID-19 and to apprehend its visibility through different kinds of indicators, such as citations and AAS.

Although the number of articles registered (2,703) is below that produced by the five most productive countries on this theme in the same period – USA (21,975), China, (9,766), UK (9,265), Italy (7,958), and India (6,146) – (Ahmed et al., 2021), yet Brazil stands out as the largest producer in research on COVID-19 in Latin America, followed by Mexico, Colombia, Chile, and Argentina (Costa et al., 2021; Espinosa et al., 2020).

The number of original investigations found in this study (53.9%) was similar to what Ahmed et al. (2021) found (57.8%) with a larger sample of documents and countries. Most of the articles were published during the second half of 2020, a relatively short time considering the lack of available data on the disease – since the first cases of COVID-19 in China were reported on December 31, 2019 – and the lack of investments in science and technology faced by Brazilian scientists.

According to De Negri et al. (2020), the inexpressive investment in science and technology, when compared to other countries, and the lack of a Brazilian strategy for research and innovation are facts that hinder the production of initiatives that face the crisis in a skillful way. In 2020, only two bids for research funding on the new coronavirus were launched by the Brazilian federal government in the amount of approximately US$ 10 million, with results released only in July of that year. In terms of comparison, only in the USA, more than US$ 6 billion was allocated exclusively to COVID-19 research. Still, we found that 623 studies (23,0%) reported some source of funding.

Even with severe budget cuts to the Brazilian universities experienced in recent years (Oliveira, 2020), the University of São Paulo still managed to stand out as the most productive institution, appearing among the 20 organizations that most published studies on the new coronavirus in the world (Bernardes & Dorado, 2020). Likewise, the most productive authors/co-authors found in this study were researchers from the University of São Paulo, a result also found by other studies that pointed out researchers from this university among the 10 most productive about COVID-19 (Gregorio-Chaviano, Limaymanta & López -Mesa, 2020; Torres Pascual & Torrell-Vallespín, 2020).

This study also shows the collaboration between scientists from Brazil and others from 133 countries for developing research on COVID-19 in multicenter studies, such as the evolution and spread of the disease (Candido et al., 2020), pharmacodynamics (Cavalcanti et al., 2020; Sterne et al., 2020) among others. This demonstrates a significant ability to articulate collaborations, in such a short time, with researchers from several countries who share a common objective, which is to mitigate the pandemic effects on the global population.

Among the research areas identified, "Public Environmental Occupational Health" (12.4%) and "General Internal Medicine" (11.4%) stood out, showing a focus on problems in the care and treatment of patients, as well as concerns about aspects of the environment and occupational health of professionals working on the front lines of this disease combat, being both multidisciplinary areas.

The results showed that English (88.1%) was the predominant language of publication. This finding reveals the concern to disseminate their studies through a language that is accessible to a greater number of researchers, which allows a wide dissemination of information. Moreover, articles in English receive more citations than those published in other languages, as higher visibility may indirectly impact the scientific community (Di Bitetti & Ferreras, 2016).

The 2,703 articles analyzed in this study received 10,190 citations, with an average of 3.77 citations per document, a value lower than that found for other countries, such as USA (8.84), China (25.62), UK (9.18), Italy (9.00), and Spain (6.43), but similar to that found for India (3.47) (Ahmed et al., 2020).

The publication of 90.6% of articles in international journals, mainly of great visibility and impact, demonstrated a positive trend in terms of research quality, in addition to the interest of Brazilian researchers in journals with the most varied scopes. In contrast, the nine journals with the largest number of publications on COVID-19 were of Brazilian origin, a result that differs from the study by Figueredo et al. (2020), who found only three Brazilian journals out of 10 with the largest number of articles published on COVID-19.

In the past year, many studies on the COVID-19 pandemic were published in scientific journals and disseminated on social media platforms, providing information and guidance on the epidemiology and characteristics of the disease, on the clinical management of patients and the development of ways of treatment and vaccines, all in real time not only for scientists and health professionals, but also for academics, mediatic, and the general public. Traditionally, articles considered to be of high relevance can be identified by counting citations received; however, this method is not always feasible to identify the impact of a publication, especially in a rapidly spreading pandemic, such as COVID-19.

When measuring the online attention received by the 100 Brazilian articles with the highest number of citations, it was found that 78.6% of all mentions on Twitter (37,965/40,062) were from members of general public, 10.4% from scientists, and 8.7% from professionals (doctors, other health professionals), showing that Brazilian articles on COVID-19 had more repercussions in the nonscientific public than in the scientific community.

Tornberg et al. (2020) investigated the correlation between AAS and citation count, suggesting an alignment between the interests of academics and the general public. In the present study, when comparing the 10 most cited articles with the 10 highest AAS, it was found that only the first three cited were among the 10 most mentioned on social platforms. A hypothesis to be considered about this would be to analyze the subject addressed in the article, which may arouse greater interest in the general public and in the press but is not so relevant to the scientific community. For example, the article with the highest AAS (6,889) addressed the use of Hydroxychloroquine with or without Azithromycin in patients with mild to moderate COVID-19 (Cavalcanti et al., 2020), a subject of great repercussions in Brazil and was widely explored by the media due to the recommendation of the use of this drug by nonspecialists in the health area, such as the then Brazilian and American presidents (*Like Trump...*, 2020). This analysis may represent an opportunity for future studies.

Conclusion

This study investigated aspects of the production, dissemination, and use of knowledge about COVID-19 produced by Brazilian authors, thus contributing to the assessment of the current state of science, as well as the impact of research in scientific community and social web. As COVID-19 still is a current topic and considering the time frame used, it is reasonable to infer that the documents retrieved in this research reflect the Brazilian scientific production and are relevant, as they were cited by other authors in their research and aroused the attention of the scientific community, but mainly of the general public. To understand the reasons for this phenomenon, we propose a content analysis to verify which specific topics are attracting greater interest from the general public, but not so much from the scientific community.

It is expected that the findings presented here will stimulate new studies on the Brazilian scientific production related to the current pandemic of COVID-19, as a way to demonstrate its relevance and the imperative necessity of investments in science even in times of political and economic difficulties.

References

Ahmed, S., Waqar, U., Lohana, K. & Khan, D. (2021). A bibliometric analysis of global COVID-19 research. *National Journal of Health Sciences*, 5(3), 119–126. doi: 10.21089/njhs.53.0119

Araujo, K., Silva, C., Guimarães, M., Lins, R. & Assef Neto, R. (2017). A Produção científica sobre zika em periódicos de acesso aberto. *Revista Eletrônica de Comunicação, Informação e Inovação em Saúde*, 11, 1–8. https://doi.org/10.29397/reciis.v11i0.1391

Baheti, A. D. & Bhargava, P. (2017). Altmetrics: a measure of social attention toward scientific research. *Current Problems in Diagnostic Radiology*, 46(6), 391–392.

Barros, M. (2015). Altmetrics: métricas alternativas de impacto científico com base em redes sociais. *Perspectivas em Ciência da Informação*, 20(2), 19–37.

Bastos, M. L., Tavaziva, G., Abidi, S. K., Campbell, J. R., Haraoui, L.-P., Johnston, J. C., Lan, Z., Law, S., MacLean, E., Trajman, A., Menzies, D., Benedetti, A. & Khan, F. A. (2020). Diagnostic accuracy of serological tests for Covid-19: systematic review and meta-analysis. *BMJ*, 370, Article m2516. https://doi.org/10.1136/bmj.m2516

Bernardes. J. & Dorado, M. (2020, October 29). *USP está entre as 20 instituições que mais publicam sobre Covid no mundo*. Jornal da USP. https://jornal.usp.br/ciencias/usp-esta-entre-as-20-instituicoes-que-mais-publicam-sobre-covid-no-mundo/

Boetto, E., Fantini, M. P., Gangemi, A., Golinelli, D., Greco, M., Nuzzolese, A. G., Presutti, V. & Rallo, F. (2021). Using altmetrics for detecting impactful research in quasi-zero-day time-windows: the case of COVID-19. *Scientometrics*, 126, 1189-1215.

Borba, M. G. S., Val, F. F. A., Sampaio, V. S., Alexandre, M. A. A., Melo, G. C., Brito, M., Mourão, M. P. G., Brito-Sousa, J. D., Baía-da-Silva, D., Guerra, M. V. F., Hajjar, L. A., Pinto, R. C., Balieiro, A. A. S., Pacheco, A. G. F., Santos, J. D. O., Naveca, F. G., Xavier, M. S., Siqueira, A. M., Schwarzbold, A., … Lacerda, M. V. G. (2020). Effect of high vs low doses of chloroquine diphosphate as adjunctive therapy for patients hospitalized with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. *JAMA Network Open*, 3(4), Article e208857. https://doi.org/10.1001/jamanetworkopen.2020.8857

Borrego, Á. (2014). Altmétricas para la evaluación de la investigación y el análisis de necesidades de información. *Profesional de la Información*, 23(4), 352–358.

Brainard, J. (2020). Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? *Science.* http://dx.doi.org/10.1126/science.abc7839

Candido, D. S., Claro, I. M., Jesus, J. G., Souza, W. M., Moreira, F. R. R., Dellicour, S., Mellan, T. A., Plessis, L., Pereira, R. H. M., Sales, F. C. S., Manuli, E. R., Thézé, J., Almeida, L., Menezes, M. T., Voloch, C. M., Fumagalli, M. J., Coletti, T. M., da Silva, C. A. M., Ramundo, M. S., … Faria, N. R. (2020). Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science*, 369(6508), 1255–1260. https://doi.org/10.1126/science.abd2161

Cavalcanti, A. B., Zampieri, F. G., Rosa, R. G., Azevedo, L. C. P., Veiga, V. C., Avezum, A., Damiani, L. P., Marcadenti, A., Kawano-Dourado, L., Lisboa, T., Junqueira, D. L. M., Silva, P. G. M. D. E., Tramujas, L., Abreu-Silva, E. O., Laranjeira, L. N., Soares, A. T., Echenique, L. S., Pereira, A. J., Freitas, F. G. R., ... Berwanger, O. (2020). Hydroxychloroquine with or without Azithromycin in mild-to-moderate Covid-19. *New England Journal of Medicine*, 383, 2041–2052.

Chahrour, M., Assi, S., Bejjani, M., Nasrallah, A. A., Salhab, H., Fares, M. & Khachfe, H. H. (2020). A bibliometric analysis of COVID-19 research activity: a call for increased output. *Cureus*, 12(3), Article e7357.

Chen, E., Lerman, K. & Ferrara, E. (2020). *#Covid-19: the first public coronavirus Twitter dataset.* arXiv. https://arxiv.org/pdf/2003.07372v1.pdf

Codo, A. C., Davanzo, G. G., Monteiro, L. B., de Souza, G. F., Muraro, S. P., Virgilio-da-Silva, J. V., Prodonoff, J. S., Carregari, V. C., Biagi Junior, C. A. O., Crunfli, F., Jimenez Restrepo, J. L., Vendramini, P. H., Reis-de-Oliveira, G., Bispo dos Santos, K., Toledo-Teixeira, D. A., Parise, P. L., Martini, M. C., Marques, R. E., Carmo, H. R., … Moraes-Vieira, P. M. (2020). Elevated glucose levels favor SARS-CoV-2 infection and monocyte response through a HIF-1α/Glycolysis-Dependent Axis. *Cell Metabolism*, 32(3), 437-446.e5. https://doi.org/10.1016/j.cmet.2020.07.00

Costa, J. P., Campos, A. L. S., Cintra, P. R., Greco, L. F. & Poker, J. H. (2021), The nature of rapid response to COVID-19 in Latin America: an examination of Argentina, Brazil, Chile, Colombia and Mexico. *Online Information Review*, ahead-of-print. https://doi.org/10.1108/OIR-09-2020-0391

De Negri, F., Koeller, P., Zucoloto, G. & Miranda, P. (2020). Investimento inexpressivo e falta de estratégia brasileira para pesquisa e inovação vão dificultar a saída da crise. *Boletim Rede de Pesquisa Solidária*, 6, 1–11.

Di Bitetti, M. S. & Ferreras, J. A. (2016). Publish (in English) or perish: The effect on citation rate of using languages other than English in scientific publications. *Ambio*, 46(1), 121–127. https://doi.org/10.1007/s13280-016-0820-7

Emanuel, E. J., Persad, G., Upshur, R., Thome, B., Parker, M., Glickman, A., Zhang, C., Boyle, C., Smith, M. & Phillips, J. P. (2020). Fair allocation of scarce medical resources in the time of Covid-19. *New England Journal of Medicine*, 382(21), 2049–2055. https://doi.org/10.1056/nejmsb2005114

Espinosa, I., Cuenca, V., Eissa-Garcés, A. & Sisa, I. (2020). Tackling COVID-19 Pandemic through Research in Latin America and the Caribbean: a bibliometric analysis. doi: 10.20944/preprints202011.0362.v1

Figueredo, W. N., Macêdo, T. T. S., Cardoso, G. M. P. & Fernandes, E. T. B. S. (2020). Análise bibliométrica da produção brasileira sobre a COVID-19. *Revista Baiana de Enfermagem*, 34, Article e37107.

Gregorio-Chaviano O., Limaymanta, C. H. & López-Mesa, E. K. (2020). Análisis bibliométrico de la producción científica latinoamericana sobre COVID-19. *Biomédica*, 40(Supl.2), 104–15. https://doi.org/10.7705/biomedica.5571

Haghani, M., Bliemer, M. C. J., Goerlandt, F. & Li, J. (2020). The scientific literature on Coronaviruses, COVID-19 and its associated safety-related research dimensions: a scientometric analysis and scoping review. *Safety Science*, Article 104806. https://doi.org/10.1016/j.ssci.2020.104806

Horbach, S. P. J. M. (2020). Pandemic publishing: medical journals strongly speed up their publication process for COVID-19. *Quantitative Science Studies*, 1(3), 1056–1067. https://doi.org/10.1162/qss_a_00076

Larivière, V., Shu, F. & Sugimoto, C. (2020 March, 5). *The Coronavirus (COVID-19) outbreak highlights serious deficiencies in scholarly communication*. LSE. https://blogs.lse.ac.uk/impactofsocialsciences/2020/03/05/the-coronavirus-covid-19-outbreak-highlights-serious-deficiencies-in-scholarly-communication/

Like Trump, Brazil's Jair Bolsonaro also bets big on chloroquine. (2020, May 21). The Japan Times. https://www.japantimes.co.jp/news/2020/05/21/world/brazil-jair-bolsonaro-hydroxychloroquine/

Melo, T. & Figueiredo, C. M. S. (2020). A first public dataset from Brazilian twitter and news on COVID-19 in Portuguese. *Data in Brief*, 32, Article 106179. http://dx.doi.org/10.1016/j.dib.2020.106179

Norte, D. B. (2020 March, 9). *Cortes e mais cortes: o que será da ciência e da pesquisa no Brasil?* VC S/A. https://vocesa.abril.com.br/carreira/cortes-bolsas-pesquisa-ciencia/

Oliveira, A. C. S. & Silva, E. M. (2016). Ciência aberta: dimensões para um novo fazer científico. *Informação & Informação*, 21(2), 5-39. http://dx.doi.org/10.5433/1981-8920.2016v21n2p5

Oliveira, E. (2020 September, 10). *Corte de quase R$ 1 bi para universidades federais é mantido mesmo com alteração no orçamento do MEC para 2021, dizem reitores*. G1. https://cutt.ly/hkZFt37

Oliveira, E. M. N., Carvalho, A. R. B., Silva, J. S., Sousa Neto, A. R., Moura, M. E. B. & Freitas, D. R. J. (2021). Analysis of scientific production on the new coronavirus (COVID-19): a bibliometric analysis. *Sao Paulo Medical Journal*, Epub January 15, 2021. https://doi.org/10.1590/1516-3180.2020.0449.r1.01102020

Piwowar, H. (2013). Value all research products. *Nature*, 493(7431), 159.

Priem, J., Taraborelli, D., Groth, P. & Neylon, C. (2010 October, 26). *Altmetrics: a manifesto*. Altmetrics. http://altmetrics.org/manifesto

Siemieniuk, R., Rochwerg, B., Agoritsas, T., Lamontagne, F., Leo, Y.-S., Macdonald, H., Agarwal, A., Zeng, L., Lytvyn, L., Appiah, J. A., Amin, W., Arabi, Y., Blumberg, L., Burhan, E., Bausch, F. J., Calfee, C. S., Cao, B., Cecconi, M., Chanda, D., … Vandvik, P. O. (2020). A living WHO guideline on drugs for Covid-19. *BMJ*, 370, Article m3379. https://doi.org/10.1136/bmj.m3379

Silva, V. R. F., Cheng, C., Silva, R. C. L., Marta, C. B., Garcia A. S., Vicentini, S. C. & Silva, C. R. L. (2020). Análise bibliométrica da produção científica sobre Coronavírus e Covid-19. *Saúde Coletiva (Barueri)*, 10(53), 2356–2369.

Sterne, J. A. C., Murthy, S., Diaz, J. V., Slutsky, A. S., Villar, J., Angus, D. C., Annane, D., Azevedo, L. C. P., Berwanger, O., Cavalcanti, A. B., Dequin, P.-F., Du, B., Emberson, J., Fisher, D., Giraudeau, B., Gordon, A. C., Granholm, A., Green, C., Haynes, R., … Marshall, J. C. (2020). Association between administration of systemic corticosteroids and mortality among critically ill patients with COVID-19. *JAMA*, 324(13), Article 1330. https://doi.org/10.1001/jama.2020.17023

Tomazini, B. M., Maia, I. S., Cavalcanti, A. B., Berwanger, O., Rosa, R. G., Veiga, V. C., Avezum, A., Lopes, R. D., Bueno, F. R., Silva, M. V. A. O., Baldassare, F. P., Costa, E. L. V., Moura, R. A. B., Honorato, M. O., Costa, A. N., Damiani, L. P., Lisboa, T., Kawano-Dourado, L., Zampieri, F. G., … Azevedo, L. C. P. (2020). Effect of dexamethasone on days alive and ventilator-free in patients with moderate or severe acute respiratory distress syndrome and COVID-19. *JAMA*, 324(13), Article 1307. https://doi.org/10.1001/jama.2020.17021

Tornberg, H. N., Moezinia, C., Wei, C., Bernstein, S. A., Wei, C., Al-Beyati, R., Quan, T. & Diemert, D. J. (2021). Assessing the dissemination of COVID-19 articles across social media with Altmetric and PlumX metrics: correlational study. *Journal of Medical Internet Research*, 23(1), Article e21408. https://doi.org/10.2196/21408

Torres Pascual, C. & Torrell-Vallespín, S. (2020). Análisis bibliométrico de la producción científica latinoamericana y del Caribe sobre COVID-19 en PUBMED. *Revista Cubana de Información en Ciencias de la Salud*, 31(3), Article e1600.

Torres-Salinas, D. (2020). Ritmo de crecimiento diario de la producción científica sobre Covid-19. Análisis en bases de datos y repositorios en acceso abierto. *Profesional de la Información*, 29(2), Article e290215. https://doi.org/10.3145/epi.2020.mar.15

Torres-Salinas, D., Robinson-Garcia, N. & Castillo-Valdivieso, P. A. (2020). *Open access and altmetrics in the pandemic age: forescast analysis on COVID-19 literature*. bioRxiv. https://doi.org/10.1101/2020.04.23.057307

Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J. H., Gelfand, M., Han, S., Haslam, S. A., Jetten, J., … Willer, R. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, 4(5), 460–471. https://doi.org/10.1038/s41562-020-0884-z

Williams, A. E. (2017). Altmetrics: an overview and evaluation. *Online Information Review*, 41(3), 311–317.

Xu, B., Kraemer, M. & Open COVID-19 Data Curation Group (2020). Open access epidemiological data from the COVID-19 outbreak. *The Lancet. Infectious diseases*, 20(5), Article 534. https://doi.org/10.1016/S1473-3099(20)30119-5

Zhang, L., Zhao, W., Sun, B., Huang, Y. & Glänzel, W. (2020). How scientific research reacts to international public health emergencies: a global analysis of response patterns. *Scientometrics*, 124(1), 747-773.

Co-Patents: What do trends reveal? An assessment of European innovation systems

Max Meier[1], Marta Romagnoli[2] , Julie Callaert[3], Paolo Landoni[4] and Bart Van Looy[5]

[1] max.meier@kuleuven.be
ECOOM – KU Leuven, Naamsestraat 61, 3000 Leuven (Belgium)

[2] marta.romagnoli@studenti.polito.it
Department of Management and Production Engineering, Politecnico di Torino, Turin (Italy)

[3]julie.callaert@kuleuven.be
ECOOM - KU Leuven, Naamsestraat 61, 3000 Leuven (Belgium)

[4]paolo.landoni@polito.it
Department of Management and Production Engineering, Politecnico di Torino, Turin (Italy)

[5]bart.vanlooy@kuleuven.be
MSI – FEB, KU Leuven, ECOOM – KU Leuven, Flanders Business School Naamsestraat 69, 3000 Leuven (Belgium)

Abstract

Concepts like 'innovation systems', 'open innovation' and innovation ecosystems all direct our attention to the importance of collaboration and spillovers for understanding innovation dynamics. In this paper we analyse the extent to which co-patents, i.e., patents owned jointly by two or more independent entities, signal a growing trend (1980 – 2015) and whether growth is situated alike across countries, technological fields and type of actors. We rely on EPO data for a set of 43 European countries. Our findings provide insights on two levels. First, a detailed assessment of the type of actors involved in co-patenting activities reveals a considerable amount (28%) of 'false positives', i.e., co-ownership of patents by actors being part of one and the same (consolidated) entity. Second, when analyzing the patterns resulting from deflated indicators, our findings reveal that co-patents are becoming more common albeit highly differentiated across fields, countries and type of actors. An outspoken increase is observed for science intensive fields as well as for governmental agencies and universities. While universities and public organizations tend to opt for more entrepreneurial behavior of a collaborative/sharing kind, this is much less the case for firms.

Introduction

Innovation and technological progress have since long been acknowledged as vital for the long term survival and growth of the firm (Baumol, 2002; Schumpeter, 1939) and as critical driving forces in enhancing social welfare. Numerous scholars have emphasized the importance to adopt a more comprehensive view to fully grasp the innovative capacity of (national) innovation systems (Freeman, 1987, 1994; Lundvall, 1992; Nelson, 1993; Mansfield & Lee, 1996; Mowery & Nelson, 1999; Dosi, 2000). The notion of 'innovation system' and more recently of 'innovation *eco*system' (Adner, 2017; Audretsch et al., 2019; Jacobides et al., 2018; Shipilov & Gawer, 2020; Xu et al., 2020) has gained wide acceptance amongst scholars and policy makers as a guiding framework to understand and model innovation systems on a more aggregated level. Within these models, not only a variety of actors (besides firms and entrepreneurs, also governmental and scientific actors, including universities and public research organizations) figure prominently; also the importance of collaboration and spillovers is being stressed.

From 2000 onwards, the concept of cooperative innovation has gained popularity, especially after Chesbrough (2003) coined the notion of 'open innovation'. Open innovation practices imply the "use of *purposive* inflows and outflows of knowledge to both accelerate internal innovation and expand the markets for external innovation." (Chesbrough, 2003; 2006b). Besides sourcing from outside, open innovation at the same time implies that inventive output from within the firm should have the opportunity to outflow into the market through a variety of channels (Belderbos et al., 2010; 2014). These mutual knowledge flows have become cornerstones of the modern literature on economic development (Belenzon, 2012).

Chesbrough's (2003) mentioning of purposefulness seems crucial for pinpointing such collaborative dynamics, by going beyond mere (potentially 'accidental') knowledge spillovers. As a relevant indicator of R&D collaborations, considerable attention has been devoted to investigating patents, which provide readily available information on such collaborations, including the technological nature, location, number and names of the individual inventors, as well as of the applicants, i.e., the owners of the underlying IP rights. Consequently, several attempts to capture 'open innovation' through patents have been made. Whereas a majority of recent work relies on co-inventorship to reflect the (geographic) scope of the collaborative process (i.e. Kogler et al. 2017; Morescalchi et al., 2015; Wanzenböck et al., 2014), yet another option resides in considering co-owned outcomes, i.e. co-owned patents as an indicator of joint R&D efforts (Arora et al., 2016; Walsha et al., 2016; Belderbos et al., 2014; Giuri et al., 2007; Chesbrough, 2006a; b) While it can be noted that not all collaborations will result in patents, let alone, co-owned patents, it is nonetheless correct to assume that co-patents do result from some form of collaborative efforts, involving a multitude of different actors (Belderbos et al. 2014). Actually, co-ownership also signals the intention of the entities to stay involved in further market exploitation efforts related to the underlying technology (Belderbos et al., 2014). As such they reflect 'joint' or 'open' business strategies, rather than joint inventive efforts.

In this paper, we systematically map and analyze the extent to which co-patents – patents owned jointly by two or more independent entities – indeed signal a growing trend; and we evaluate and compare growth patterns (time period 1980 – 2015) across countries, technological fields and type of actors (sectors). Before analyzing the occurrence of co-patent activity (over time and across fields, countries) we first address a risk of biased measurement of collaboration in technology when considering jointly owned patents. Numbers may become inflated due to patents that are jointly owned by different departments of one and the same organization, rather than by independent entities. This paper sets out by deflating the indicators accordingly. Next, we use the correct (deflated) indicators on technological collaboration, to evaluate patterns over time. We thereby distinguish between the variety of actors (firms, individuals, universities, governmental agencies) populating the innovation system. By analyzing the sectors engaged in co-patenting, and the occurrence of technological collaboration across technology domains and countries, we are able to shed light on whether the tendency to adopt more collaborative practices is indeed widespread across European innovation systems and/or more situated within specific parts (type of actors) of innovative ecosystems.

As such, we contribute to and complement previous work on co-patenting and more generally, on assessing trends in the involvement of multiple actors in R&D collaborations. In particular, the decline of science in corporate R&D, as recently signalled by Arora et al. (2017) is a point of interest. Arora et al. (2017) observe that large corporations are becoming more reluctant to invest into their own internal scientific capabilities while at the same time relying more on outside collaborations in this respect. The question arises whether indicators signalling co-ownership (of patents) reflect such 'division of labour' trends; is co-ownership between firms and universities and/or public research organizations becoming more prevalent recently?

Data and methodology

To investigate the presence of co-ownership in patenting, data has been retrieved from the PATSTAT.2018b database, which provides exhaustive information on patent data, i.e. patent grants and applications to the European Patent Office (EPO), along with details on the applicants, priority years, applicant countries and technology fields (EPO, 2018). In a first step, information on the applicant(s) has been used to distinguish between patent applications filed by a single applicant from jointly filed patent applications implying two or more applicants that (seemingly) represent independent entities. In a second step, this information was used to highlight co-patenting trends over time, for different actor-combinations (sectors) and technological fields. For both, we considered all patent documents with priority years between 1980 and 2015 and included a set of 43 European countries[i], rendering a total patent count of 1.558.354 patents (including both single-and co-owned).

Deflating co-patent counts

Co-patents are identified as patents with more than 1 applicant. All applicants have been harmonized according to the name harmonization algorithm developed at KU Leuven. This procedure already distinguishes between different departments within organisations (for details on the procedure, see Van Looy et al., 2006; Eurostat 2011). Closer inspection of co-patenting partners however revealed the presence of 'false positives': co-applicants that are in reality different departments, business units, subsidiaries etc. of one and the same organisation, rather than independent entities. Examples of seeming co-applicant pairs include 'NOVARTIS' patenting with 'NOVARTIS ERFINDUNGEN VERWALTUNGSGESELLSCHAFT', 'UNILEVER UK' co-patenting with 'UNILEVER NL', 'SCHLUMBERGER TECHNOLOGY co-patenting with 'SCHLUMBERGER HOLDINGS' or 'SERVICES PETROLIERS SCHLUMBERGER, and 'PHILIPS ELECTRONICS' with different country-specific subsidiaries, i.e. 'PHILIPS ELECTRONICS UK', 'PHILIPS DEUTSCHLAND' or 'PHILIPS NORDEN'. If one defines co-patenting as collaboration between truly independent entities, then the inclusion of such cases leads to an overestimation of the co-patenting phenomenon and inflated co-patenting statistics. To correct for this, we calculated Levenshtein distance (LV) measures between all co-patent applicant names. A human rater decision tree was designed, based on LV threshold values and sector allocation results (based on the sector allocation algorithm developed at KU Leuven, see Eurostat 2011). After inspection of pairs according to the designed procedure, we observe the following difference in Table 1 between the inflated and deflated co-/ single patent count, suggesting a total of 'false positives' amounting to approximately 28%.

Table 1. Number of co-patents before and after removal of false positives.

	Co-Patents	*Single owned Patents*	*Total Patents*	*Co-patent shares (%)*
Inflated count	117.866	1.440.488	1.558.354	7,56
Deflated Count	85.397	1.472.957	1.558.354	5,48

In Figure 1, we present the deflated and inflated co-patent share over time, showing the systematic overestimation of the co-patenting phenomenon. Furthermore, we observe considerable variability of the trend, albeit with seemingly more consistent increases since the 2000s, from a comparatively lower level.

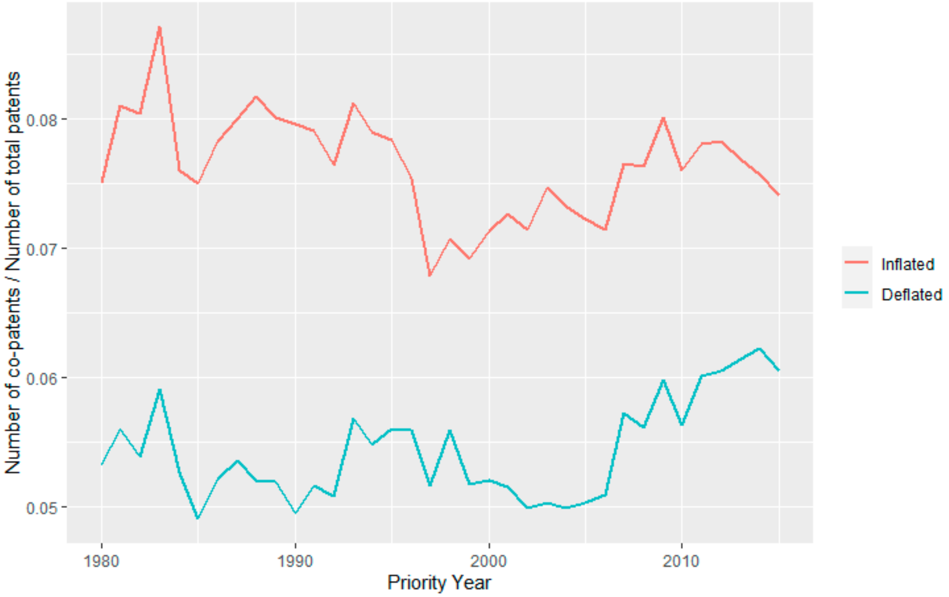


Figure 1. Comparison of inflated vs. deflated co-patent application shares.

Identification of co-patent trends

To provide a differentiated view of co-patent trends, we utilize the *deflated* co-patent count and explore the resulting patterns across fields, countries and type of actors. For technology fields, we rely on the technology classification provided by Fraunhofer (FhG), which assigns IPC patent classes to 35 broad technology fields, such as Telecommunications, Semiconductors, Biotechnology or Pharmaceuticals. Furthermore, we exploit the country and sector information related to the patent applicants. For each patent, we define whether the patent is (co-)owned by companies, individuals, universities and/or governmental institutions, whereas all other categories have been aggregated to "other" sectors (including hospitals and ambiguous legal entities). In cases of multiple assignments (i.e., one patent belonging to multiple technology fields, to multiple sectors, and/or to applicants in different countries), we made use of the full patent count, i.e., counting the patent with 1 for every technology field/ applicant country/ sector. When considering sector combinations at the individual patent level, we limit our analysis to cases of maximum two different sectors (i.e., company-university), whereas more complex constellations (i.e., company-university-government) are assigned to "other". We then explore some basic trends, and additionally perform an analysis of covariance (ANCOVA) to identify if and to what extent, co-patenting is country- and technology-field specific. The ANCOVA allows to evaluate whether the means of a dependent variable are equal across different levels of categorical independent variables, while statistically controlling for the effects of other continuous variables (covariates). Applying a minimum threshold of 105 out of 1260 (35 *FhG_class**36 *Prio_year)* possible occurrences with at least 1 co-patent, we limit the analysis to a subset of 25 applicant countries (*Applt_ctry*) and all 35 technology fields (*FhG_class*) as categorical variables and estimate their effect on the share of co-patents (*Co-pat_share*), while controlling for the application year (*Prio_year*).

Analysis and results

Observing the co-patenting trends over time by utilizing the deflated co-patent counts, reveals a highly differentiated picture of co-patenting as a phenomenon. Figure 2 shows a u-shaped evolution of co-patenting, with a downward trend until the mid-1990s and a considerable upward trend thereafter.

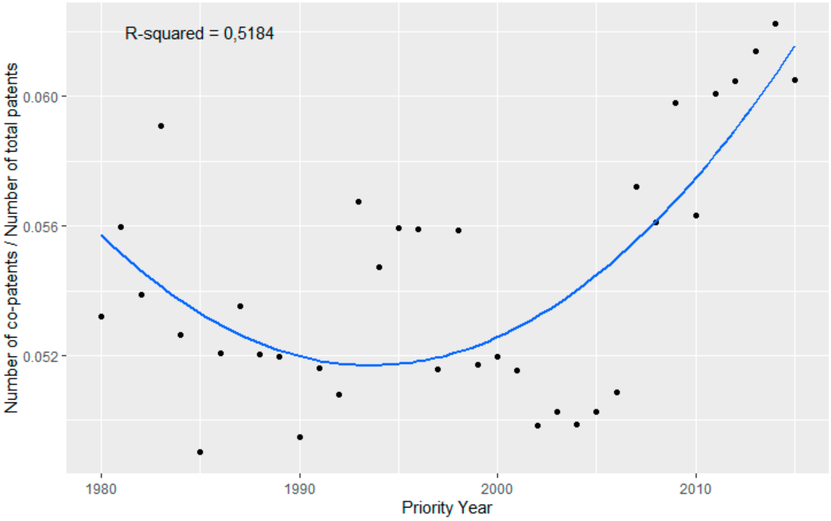


Figure 2. Evolution of deflated co-patent share (total share).

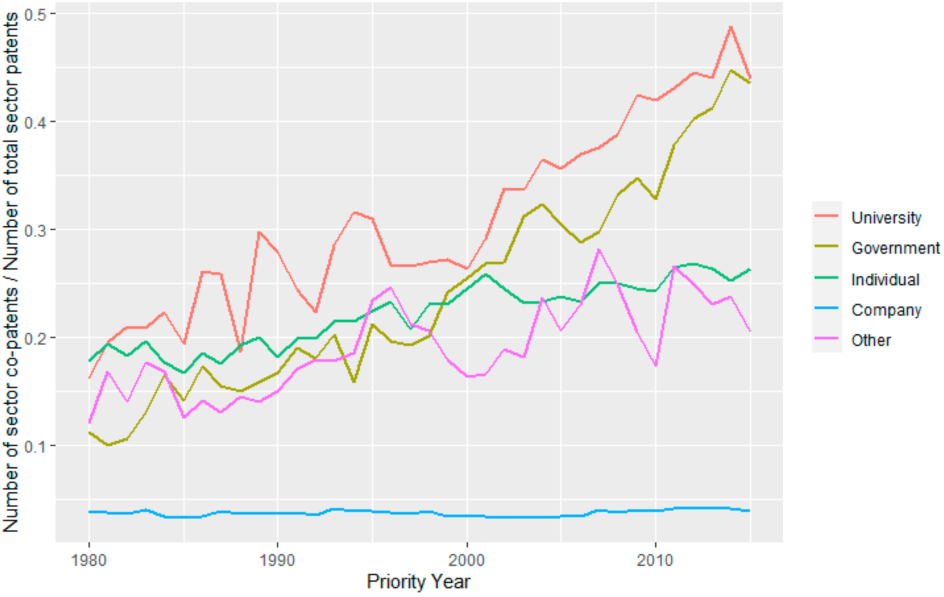


Figure 3. Comparison of deflated co-patent application shares within sectors.

Figure 3 provides an overview of the co-patent application shares within different sectors over time. While co-patenting for companies (as a share of total patents filed by companies) remains stable at a comparatively low level (<5%), co-patents filed by universities and governments (as a share of total patents filed by universities and governments respectively) display a sharp increase, growing from around 16% and 11% in 1980 to around 44% in 2015. This observation is furthermore confirmed in Table 2, showing a strongly growing trend in terms of co-patent counts and shares for all sectors, while a decreasing, albeit non-significant downward trend in shares for companies.

Table 2. Time evolution of deflated co-patent application shares/ counts within sectors.

	Co-patent	*Application Year*
University	Share (%)	0.937**
	Count	0.911**
Government	Share (%)	0.965**
	Count	0.924**
Individual	Share (%)	0.930**
	Count	0.736**
Company	Share (%)	0.267
	Count	0.969**
Others	Share (%)	0.774**
	Count	0.850**

** $p \leq 0.01$

Going beyond the single entity view, in Figure 4 and Figure 5 we disaggregated the results to investigate co-patenting sector pairs. While the numbers of co-patents have grown considerably over time, this growth seems to have been largely driven by sector pairs in which universities and governmental agencies – rather than companies – figure as co-owners.

Observing the relative share of co-patent sector pairs (among the total co-patents), the combination company-company has steadily decreased from around 41% in 1980 to 29% in 2015. Similarly decreasing trends can be observed for co-patents filed by individuals, and at a lower level for company-individuals. On the other hand, combinations involving universities or governments have all increased, most prominently for government-university co-patents, with a visible surge especially after around the year 2000 (from ~2% to around 13% in 2015).

Following this, in Figure 6, we observe a decreasing *share* from ~91% of private-private combinations in 1980 to ~52% in 2015, whereas public-public combinations have increased their share from <1% to ~20%. Similarly, public-private combinations have increased from around 7% to around 22%. Taken together, these trends suggest that the relative increase in co-patenting is predominantly driven by actor combinations that involve public domain actors.

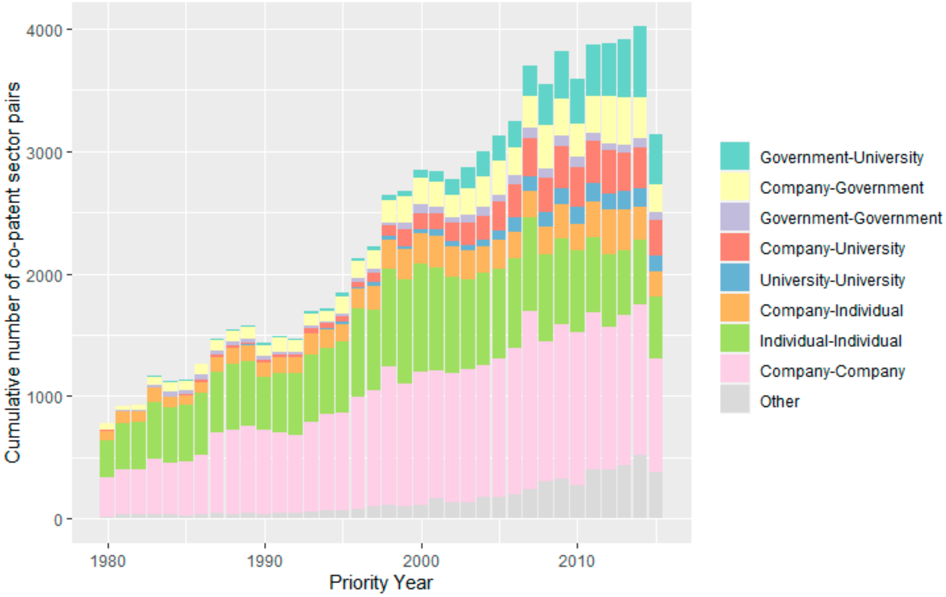


Figure 4. Distribution of deflated co-patent sector pairs.

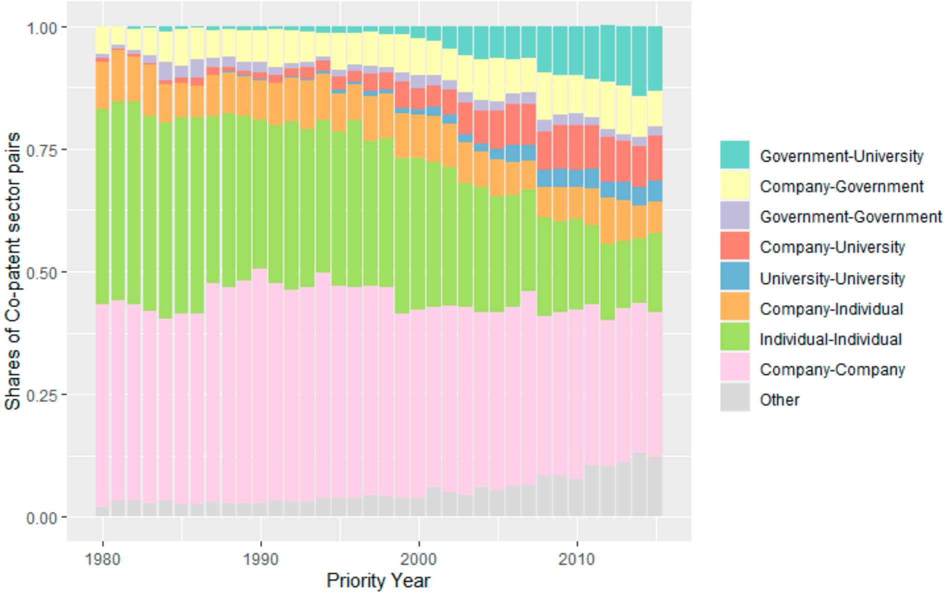


Figure 5. Relative distribution of deflated co-patent sector pairs.

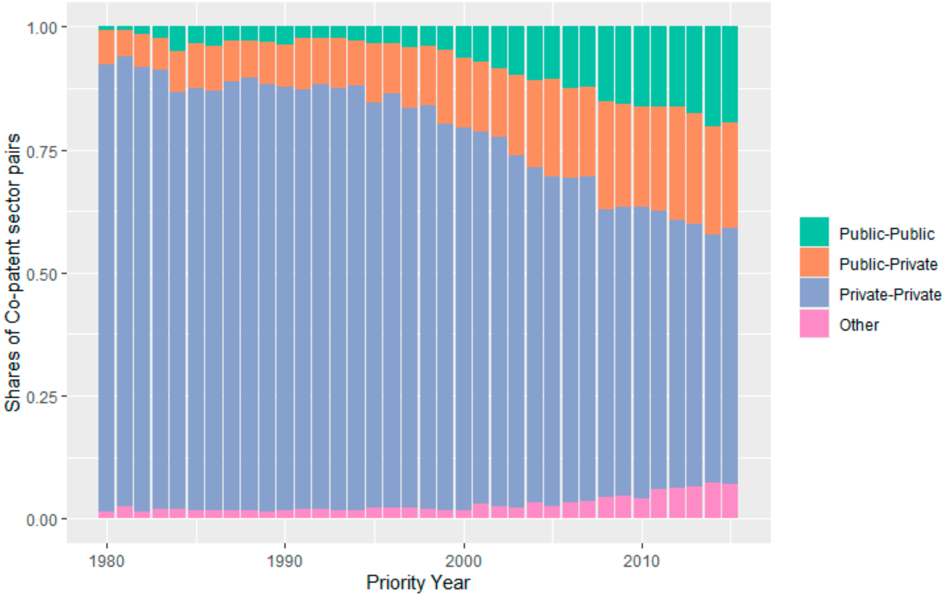


Figure 6. Relative distribution of deflated co-patent sector pairs (public-private).

Turning to the technology fields, an overview of the share of co-patents in each field is provided in Figure 7. The technology fields with the lowest co-patenting rates (below or around 4%) are Telecommunications (3), Digital Communication (4) and Handling in the Mechanical Engineering area (25). The median co-patent share for all technology fields is ~6%.

In contrast to that, we find comparatively high co-patenting shares for four technology fields ranging from 12% to 15%, involving Micro-structure and nano-technology (22), Analysis of biological materials (11), Biotechnology (15) and Pharmaceuticals (16). These results indeed suggest a high field-heterogeneity of the co-patenting phenomena. In particular, the shares of co-patenting seem to be higher for science-intensive fields.

Observing the relative growth dynamics over time in Figure 8, reveals further interesting insights. First, we note considerable year-to-year variability, in particular in the 1980s, where some fields have relatively low patenting rates (i.e., only 18 patents in the field Analysis, Measurement and Control Technology (7) in 1981, or in the 1990s after the first introduction of Micro-structure and nano-technology patents with a yearly patent count of 19 to 52 between 1992 and 1996), leading to a rather large variation of co-patent shares at the beginning of the observed timeframe. Second, we also observe steady growth patterns, in particular for Pharmaceuticals, which initially had a relatively low co-patenting share. Overall, we observe co-patenting rates reaching 10 to 15% by the end of the 1990s in those science-intensive four fields as mentioned above. In addition, after the year 2000, we observe another surge in co-patenting shares, closer to or exceeding 20% until 2015, clearly distinguishing these fields from all others.

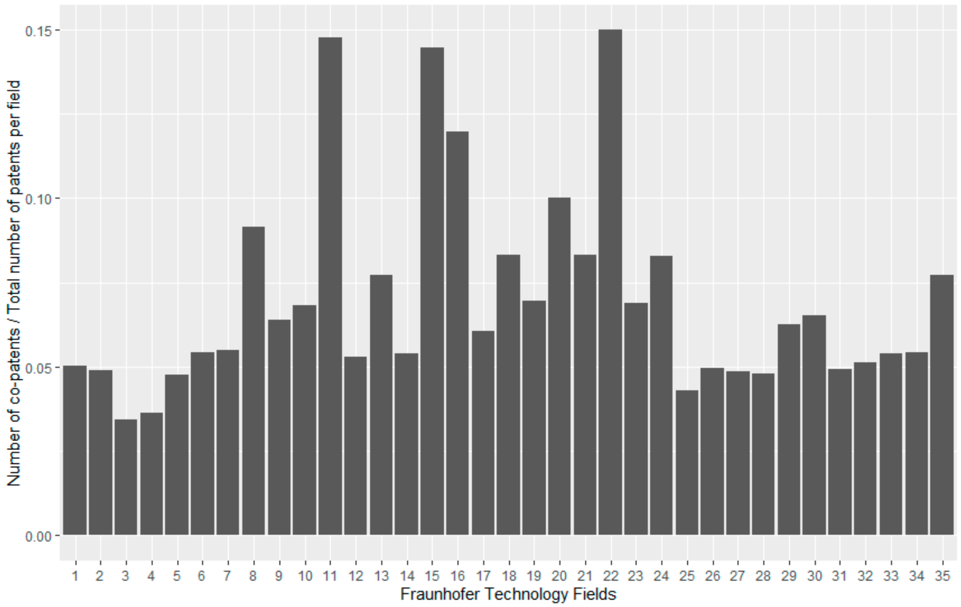


Figure 7. Comparison of co-patent applications by technology fields.

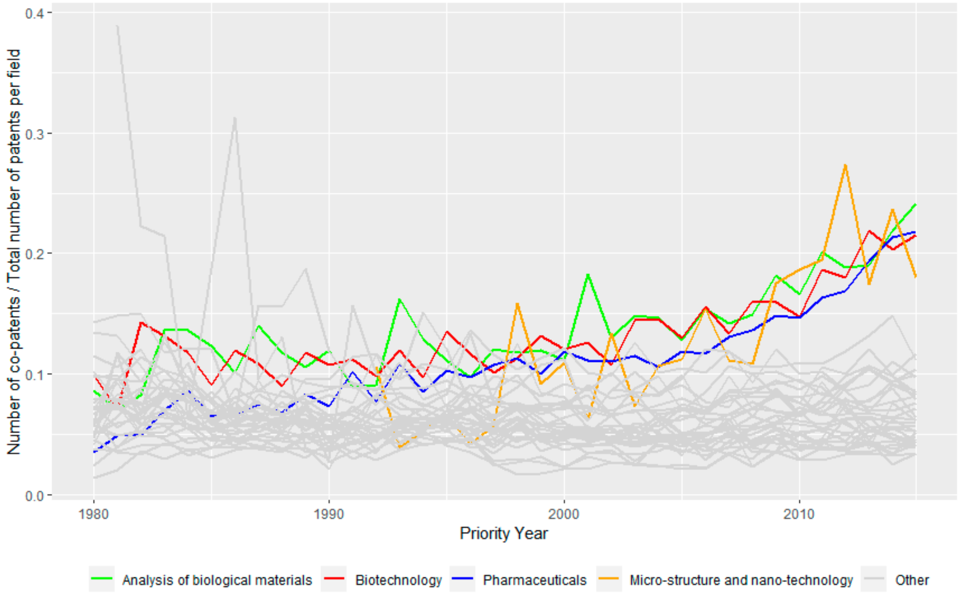


Figure 8. Comparison of co-patent applications by technology fields over time.

In addition, we aimed to identify if and to what extent, co-patenting is also country specific (besides domain/field specific). As can be seen in the ANCOVA output in Table 3, co-patenting varies significantly across countries and technological fields. In particular, we find a strong country-dependence, which explains by far the largest part of the model fit as compared to individual influences of the application year and the technology field.

Table 3. ANCOVA.

Dependent variable: *Co-pat_share*

Origin	*Sum of Squares (Type III)*	*df*	*Root mean squared*	*F*	*Sign.*
Corrected model	571,140	863	0,662	38,474	0,000
Intercept	159,943	1	159,943	9298,154	0,000
Prio_year	0,565	1	0, 565	32,833	0,000
*Prio_year * Prio_year*	0,054	1	0, 054	3,146	0,076
FhG_class	36,158	24	1,063	61,824	0,000
Applt_ctry	371,646	34	15,485	901,121	0,000
*FhG_class * Applt_ctry*	84,290	803	0,105	6,102	0,000
Error	277,445	16129	0,017		
Total	1366,621	16993			
Correct Total	848,585	16992			

R-squared = 0.673 (R-squared adjusted = 0.656)

Discussion and conclusion

In this paper, we aim to contribute novel insights to the phenomenon of and literature on R&D collaborations and its role within (national) innovation systems. As such, we systematically mapped and analyzed the extent to which co-patents, i.e. patents owned jointly by two or more independent entities signal a growing trend, and whether and to what degree this growth is differentiated across countries, technological fields and type of actors. By relying on patent applications to the EPO for a set of 43 European countries, over the time period 1980-2015, we provide several contributions.

First, from a methodological point of view, we highlight the need to carefully consider data consolidation concerns, when relying on existing information on co-owned patents in patent databases. By defining co-patents strictly as those jointly owned by independent entities, and thereby excluding joint applications of different departments or subsidiaries from one and the same consolidated entity, our results reveal a considerable amount of 28% 'false positives' for our dataset. Not only does this represent an overestimation of the 'true' phenomenon of co-patenting; trends derived from 'deflated' figures present themselves as opposite to the ones obtained from the 'raw' data/indicator (see Figure 1).

Second, when analyzing trends based on the deflated indicator, we indeed observe a growing importance of co-patenting as a tangible output of R&D collaborations. At the same time this trend is highly differentiated across fields, countries and type of actors. While increasing trends are particularly present for some science-intensive fields (such as pharmaceuticals, biotechnology or nanotechnology) where almost 20% of the patent applications are co-patents by the end of the observed timeframe, most other fields remain situated at stable co-patent shares, below 10%. In terms of countries, we observe even stronger differences, signaling outspoken differences in the configuration and collaborative practices between the respective innovation systems. In particular, the variety of actors populating the innovation system seems to be a driving force, as the relative increase in co-patenting is predominantly stemming from actor combinations that involve universities and/ or governments, ergo entities of the public domain.

These findings resonate with earlier contributions focusing on the presence and impact of co-patenting. Belderbos et al. (2014) advance a distinction between value creation (joint development) and value appropriation (joint exploitation) in R&D collaborations. As such, co-patenting – oriented towards appropriation – is approached cautiously by companies due to looming downward appropriation effects resulting from shared intellectual property (in particular with companies active in similar industries). Consequently, the observation that firms do not increase their (relative) level of engagement in co-patenting practices comes hardly as a surprise. In line with the observations advanced by Arora et al. (2017), we do observe an increasing trend whereby firms co-patent more with universities and public research organizations. Equally interesting to notice – as well as to further analyze and substantiate in terms of antecedents – is the behavior of public actors regarding co-patenting. Our findings reveal clearly that universities and governmental agencies not only become more active in terms of technology development; they tend to combine this entrepreneurial stance with ownership practices of a more collaborative nature. We hope our contribution inspires future research, both in terms of antecedents and consequences of these differentiated practices.

References

Adner, R. (2017). Ecosystem as structure: An actionable construct for strategy. *Journal of Management*, 43(1),39–58.

Arora, A., Athreye, S. & Huang, C. (2016). The paradox of openness revisited: Collaborative innovation and patenting by UK innovators. *Research Policy*, 45(7), 1352–1361.

Arora, A., Belenzon, S. & Patacconi, A. (2017). The decline of science in corporate R&D. *Strategic Management Journal*, 39(1), 3–32.

Audretsch, D. B, Cunningham, J.A., Kuratko, D.F., Lehmann, E.E. &Menter, M. (2019). Entrepreneurial ecosystems: Economic, technological, and societal impacts. *The Journal of Technology Transfer*, 44(2), 313–325.

Baumol, W.J. (2002). *The Free-Market Innovation Machine: Analyzing the Growth Miracle of Capitalism.* Princeton: Princeton University Press.

Belderbos, R., Cassiman, B., Faems, D., Leten, B. & Van Looy, B. (2014). Co-ownership of intellectual property: Exploring the value-appropriation and value-creation implications of co-patenting with different partners. *Research Policy*, 43(5), 841-852.

Belderbos, R., Faems, D., Leten, B. & Van Looy, B. (2010). Technological activities and their impact on the financial performance of the firm: exploitation and exploration within and between firms. *Journal of Product Innovation Management*, 27, 869-882.

Belenzon, S. (2012). Cumulative innovation and market value. *The Economic Journal*, 122, 265-285.

Chesbrough, H. (2003). The era of open innovation. *MIT Sloan Management Review*, 44, 35-41.

Chesbrough, H. (2006a). *Open Business Models: How to Thrive in the New Innovation Landscape*. Harvard Business Press.

Chesbrough, H. (2006b). Open Innovation: A new paradigm for understanding industrial innovation. In H. Chesbrough, W. Vanhaverbeke & J. West, (Eds.), *Open Innovation: Researching a new paradigm* (pp.1-12). Oxford: Oxford University Press.

Dosi, G. (2000). *Innovation, Organization and Economic Dynamics.* Cheltenham: Edward Elgar Publishers.

European Patent Office. (2018). PATSTAT database, Autumn 2018.

Eurostat. (2011). Patent Statistics at Eurostat: Methods for Regionalization, Sector Allocation and Name Harmonisation.

Freeman, C. (1987). *Technology policy and economic performance*. London: Pinter.

Freeman, C. (1994). The economics of technical change. *Cambridge Journal of Economics*, 18, 463-514.

Giuri, P., Mariani, M., Brusoni, S., Crespi, G., Francoz, D., Gambardella, A., Garcia-Fontes, W., Geuna, A., Gonzales, R., Harhoff, D., Hoisl, K., Le Bas, C., Luzzi, A., Magazzini, L., Nesta, L., Nomaler, Ö., Palomeras, N., Patel, P., Romanelli, M. & Verspagen, B. (2007). Inventors and invention processes in Europe: Results from the PatVal-EU survey. *Research Policy*, 36(8), 1107–1127.

Jacobides, M., Cennamo, C. & Gawer, A. (2018). Towards a theory of ecosystems. *Strategic Management Journal,* 39(8), 2255–2276.

Kogler, D. F., Essletzbichler, J. & Rigby, D. L. (2017). The evolution of specialization in the EU15 knowledge space. *Journal of Economic Geography*, 17(2), 345–373.

Lundvall, B.A. (1992). *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning*. London: Pinter Publishers.

Mansfield, E. & Lee, J.Y. (1996). The modern university: contributor to industrial innovation and recipient of industrial support. *Research Policy*, 25, 1047-1058.

Morescalchi, A., Pammolli, F., Penner, O., Petersen, A. M. & Riccaboni, M. (2015). The evolution of networks of innovators within and across borders: Evidence from patent data. *Research Policy*, 44(3), 651–668.

Mowery, D.C., & Nelson, R.R. (1999). *Sources of Industrial Leadership.* Cambridge: Cambridge University Press.

Nelson, R.R. (1993). *National Innovation Systems: A Comparative Analysis.* New York: Oxford University Press.

Schmoch, U. (2008). Concept of a Technology Classification for Country Comparisons. Final Report to the World Intellectual Property Organisation (WIPO).

Schumpeter, J.A. (1939). *Business Cycles: A theoretical, historical and statistical analysis of the Capitalist Process*. New York: McGraw-Hill.

Van Looy, B., Du Plessis, M. & Magerman, T. (2006). Data production methods for harmonized patent indicators: assignee section allocation. *KU Leuven Working Paper*, No. MSI 0606.

Walsh, J. P., Lee, Y.-N. & Nagaoka, S. (2016). Openness and innovation in the US: Collaboration form, idea generation and implementation. *Research Policy*, 45(8), 1660–1671.

Wanzenböck, I., Scherngell, T. & Brenner, T. (2014). Embeddedness of regions in European knowledge networks: A comparative analysis of inter-regional R&D collaborations, co-patents and co-publications. *The Annals of Regional Science*, 53(2), 337–368.

Xu, G., Hu, W., Qiao, Y. & Zhou, Y. (2020). Mapping an innovation ecosystem using network clustering and community identification: A multi-layered framework. *Scientometrics,* 124, 2057–2081.

[i] As defined by United Nations Statistics, excluding dependencies and the Holy See.

Locally global. Understanding the structural evolution of a bi-lingual national research landscape.

Matias Federico Milia[1,2], Ariadna Nebot Giralt[2] and Rigas Arvanitis[2,3,4]

[1] *matias@milia.net*
FLACSO-México, Carretera al Ajusco 377, Colonia Héroes de Padierna, Tlalpan, Ciudad de México (México)
[2] Global Research Institute of Paris, Université de Paris, 85 boulevard Saint-Germain, 75006 Paris (France)
[3] Centre Population et Développement, Université de Paris, Campus Saint-Germain, 45 rue des Saints-Pères, 75006, Paris (France)
[4] Institut de Recherche pour le Développement, 54 Boulevard Raspail, 75006 Paris (France)

Abstract

Research institutions organize their scientific activities in an increasingly diverse landscape. In matters of global interest, research relies on an ever-more cross-disciplinary background, which reveals intriguing questions concerning the local dynamics vs. global audiences. This paper proposes new methodological tools to assess, from a strategic perspective, the evolution of a given research landscape. It relies on the Global Research Institute of Paris's recent experience, a new interdisciplinary Institute that focuses on globalization topics beyond the usual economic meaning. The Institute relies on a broad and diverse set of research units of the Université de Paris and relates to the broad landscape of social sciences in France. This article charts the evolution of French authors' scientific publications on the Institute's thematic interests in French and English. It focuses on the structural features of the debate, namely the volume and the underlying semantic structure. The paper offers significant evidence to understand knowledge circulation dynamics and links that non-speaking countries' scientific literature builds with the English one.

Introduction

Research institutions build their research agendas in an ever-more international, dynamic, and diverse landscape. Many of these efforts are increasingly cross-disciplinary, either in collaboration or competition. Various institutions are motivated to inform their activities through the analysis of available bibliometric data. Understanding a specific research landscape is critical to steer global and multidisciplinary research agendas (Wallace & Ràfols, 2015, 2018). Different approaches tend to emerge in research areas with an international scope, gathering varied research interests from diverse origins and disciplinary backgrounds. One way of observing these complex groupings is to assess emerging trends over time (Ràfols, Porter, & Leydesdorff, 2010; Zitt & Bassecoulard, 1994). It is possible thus to describe the emerging tendencies and follow their evolution.

However, such an interpretative approach may reveal not so much the strategic choices of authors and affiliated institutions but rather the limitations experienced by researchers in their academic environment (Cruz-Castro & Sanz-Menéndez, 2018). Institutional structures tend to reflect the history of the scientific endeavor (Trow 1999, p.4). These questions are important when considering the emergence and circulation of new thematics. Local interests may differ from those appearing in the global arenas, but research in other languages tends to be overlooked in English literature (Chi 2012). How global transformations of research affect local scientific choices is not only a challenging and interesting question; it is also central to understand the growing globalization processes (Arvanitis & O'Brien, 2019).

We present the work done at the Global Research Institute of Paris (GRIP), a new French Institute that wanted to map, from an institutional perspective, research areas with a global scope. In particular, GRIP wanted to investigate the possible distance between its own strategic choices and those found in the French and international literature.

GRIP is an interdisciplinary institute focussing on globalization beyond its usual economic dimensions. GRIP builds upon the broad and diverse social science landscape inside the perimeter of the Université de Paris. Its interpretation of the concept of "Global Research" bounds to a vast array of research units scattered over Paris and relates to the broad landscape of social sciences in France. The multilingual publishing practices toward diverse audiences in social science research become a privileged standpoint of analysis (Kulczycki et al., 2020).

GRIP chose to concentrate on three thematic areas: Global Citizenship; Circulations; and Technologies, Market Logics and Vulnerabilities. These topics have been chosen through exchanges among academics located in the University, with no previous strategic analysis. It was decided to confront these strategic choices with the topics appearing in the international 'mainstream' publications and native-french publication repositories. The aim was to see if the Institute's topics are different or coherent with the available literature.

Methods and materials

This paper charts published research related to the same notion of Global Research that inspired GRIP's development and in two different languages (French and English). Specific queries were built to capture documents involving each of the Institute's thematic areas. As it remains crucial to situate these interpretations, this research focuses exclusively on French authors. To represent the GRIP's scientific objects and interests, we shared our research queries with a sample of researchers involved in the Institute's thematic areas. In this way, we focused on defining a precise query for each thematic area. After collecting and examining our first harvest of documents, the feedback from these same researchers became an input to fine-tune our queries.

As a result, six different queries for information retrieval were written, one for each research axis in both of the chosen sources. For English documents, information was retrieved from the Web of Science database. Items written in French were accessed from the HAL-SHS collection using the API made available by the platform (CCSD, 2021). HAL-SHS is a mandatory repository for French researchers that has been poorly explored and remains highly relevant since regional databases are an essential source to study social sciences' research output (Engels et al., 2012). In both cases, the fields we retrieved were the titles, abstracts, keywords, authors, and publication years. In WoS, the search strategy was based on the documents' topics. In HAL-SHS, full texts were available and searched; we excluded doctoral thesis and other teaching-related documents. We refined and improved our search strategy by consecutive iterations after analyzing the results in a continuous dialog with the researchers, experts in the field. As a result, we established the following query lines for each language, as shown in Table 1.

In this process, we gathered 1,798 English documents from the Web of Science (WoS) and 4,545 French documents from the HAL-SHS, both between 1980 and 2020. The query aimed at articles, books, book chapters, and research communications in both sources. We processed the two resulting collections with the help of the CorText platform using NLP techniques to obtain a fine-grain meaningful set of 750 terms describing the main semantic features of both collections. These lists of terms were manually curated[i] to eliminate noise and shallow terms in both French and English.

A *Period Detector* script was used to capture the semantic structure. We analyzed the frequency distribution of significant terms over the years and computed the optimal partitions in subsequent periods. We selected only the words with a minimum frequency of eight, 376 terms in French and 351 in English. Periods were computed using the gap statistic method (Tibshirani et al., 2001). As a result, we could improve the stability of differences among subsequent years. In this way, we gain access to the underlying knowledge aggregation process reflected in these documents.

Table 1. Queries used to represent each thematic axis in WoS (English) and HAL-SHS (French).

Web of Science (English)	**HAL – SHS** (French)
Axis 1: *Global Citizenships*	
TS = ((((city OR cities) NEAR/2 (irregular migration)) OR ((city OR cities) NEAR/2 (migrant workers)) OR ((city OR cities) NEAR/2 (transnational labor))) OR ((((sanctuary city) OR (sanctuary cities)) OR (((municipal* OR local) NEAR/2 Government*) NEAR (*migration policy)) OR (asylum NEAR/2 (city OR cities)))) OR (((global OR world) NEAR/1 (city OR cities OR megacity OR megacities OR metropolis OR megalopolis))) OR (((((city OR cities OR urban) NEAR/2 (practice*)) OR ((city OR cities OR urban) NEAR/2 (ways of living))) AND (digital OR digitalization)))) AND CU=(FRANCE)	q=text_fulltext (((ville OR villes) AND (cosmopolitisme OR nouvelle technologie OR globalisation OR Ségrégation OR ("intrusion exclusion" ~2))) OR ("droit à la ville") OR (("ville sanctuaire" ~2) OR ("ville accueillante")))
Axis 2: *Circulations*	
TS = (((CSR OR (Corporate Social Responsibility)) AND (global* OR international* OR transnational*)) OR (((region* OR territor* OR local*) NEAR (development)) AND (agricultur* OR agribusiness OR industr* OR producers OR production OR productive OR productivity) AND (value chains OR export* OR liberalization OR (non NEAR tariffs) OR FDI OR (foreign investment) OR openness OR transnational OR international OR global* OR trade OR (economic cooperation))) OR (multilingualism) OR (((maker movement) OR (inventive activity) OR (history NEAR (technolog* OR innovation)) OR (thing turn))) OR (((remittances OR transnationalism) AND ((non-economic) OR political OR cultural OR scientific OR scientist* OR artist* OR (higher education) OR businessmen OR academic OR commercial))) OR (((global OR transnational OR international OR cross-national) AND (diffusion OR circulation OR flow* OR networks OR isomorphism) AND (((institutions OR institutional) NEAR model) OR (organizational practices) OR education OR teaching OR art OR culture OR music OR stantard* OR standardization OR certification)))) AND CU=(FRANCE)	q=text_fulltext (("mondialisation par le bas") OR (("nouvelle technologie" OR "nouvelles technologies") OR ("réseaux sociaux" AND "processus identitaire")) OR ("robotique humanoïde") OR ((circulation OR globalisation OR mondialisation) AND (management responsable OR RSE OR "Responsabilité Sociale des Entreprises")) OR (métropolisation OR "Aménagement du territoire" OR "dévelopement régional") OR (plurilinguisme) OR (transfert OR musique OR norme OR idée* OR savoirs OR argent OR politiques))
Axis 3: *Technologies, Market Logics and Vulnerabilities*	
TS = ((((STS OR "science studies" OR (science NEAR/2 technology NEAR/2 society)) AND ("global south" OR periphery OR "latin america" OR "latin-america" OR postcolonial OR Asia OR Africa OR India))) OR ((medicine OR health) AND (globalization OR transnational)) OR (((("digital technologies" OR "information and communication technologies") AND health) OR mHealth OR "m-Health" OR "eHealth" OR "digital health") AND (gender OR women)) OR ((anthropocene OR globalization OR globalizing OR transnational) AND (pollution OR environment OR environmental)) OR ((("science" OR "scientific" OR "research") AND ("crowdsourcing")) OR ("crowd science") OR ("participative research") OR ("participatory research") OR ("citizen science"))) AND PY=(2009-2020) AND CU=(FRANCE)	q=text_fulltext (((STS OR "science studies" OR "science and technology studies" OR technoscience) AND (Sud OR Suds)) OR ((globalisation OR mondialisation) AND (medicines OR santé)) OR (((santé AND (numérique OR mobile)) OR (mhealth OR mSanté)) AND (maternelle OR sex* OR genre*)) OR ((globalisation OR mondialisation) AND (anthropocène OR environnement OR pollution)) OR ("Sciences participatives" OR "science participative" OR ("sciences citoyennes") OR ("recherche participative")))

Findings

This analysis describes the themes related to the scientific interests of the GRIP in French publications in two different languages. We discovered that within these issues, as defined above, more documents were published before 2005 in French (393 documents) than in English (115 docs), and publications grow earlier and faster in French than in English (Figure 1). When comparing both curves, there is a significant time lag between French and English documents. The volume of publications around these issues picks up in French between 2002 and 2003, but only six years later in English, between 2008 and 2009. The French language growth seems to feed

English-language papers, but with a significant time lag. Finally, we observe that the number of published English-language documents stagnates from 2018 and on.

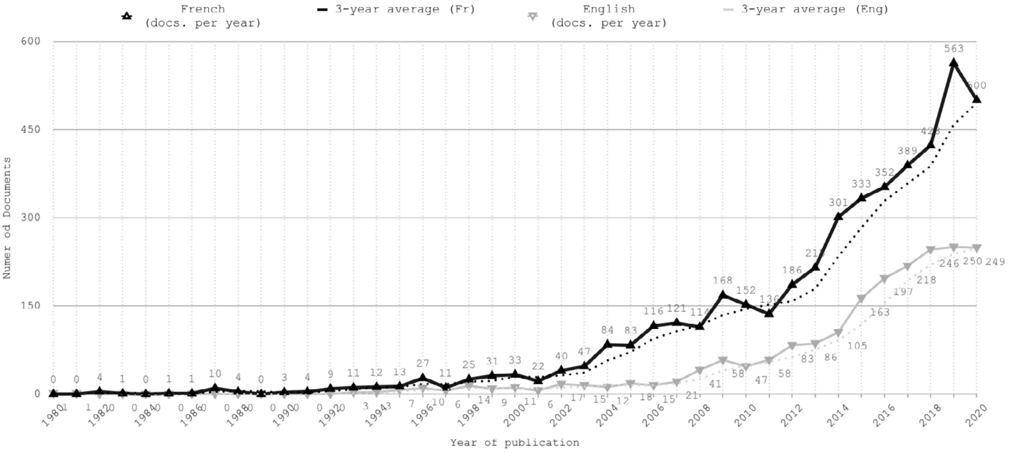


Figure 1. Number of documents representing the Institute's thematic interests in English and French.

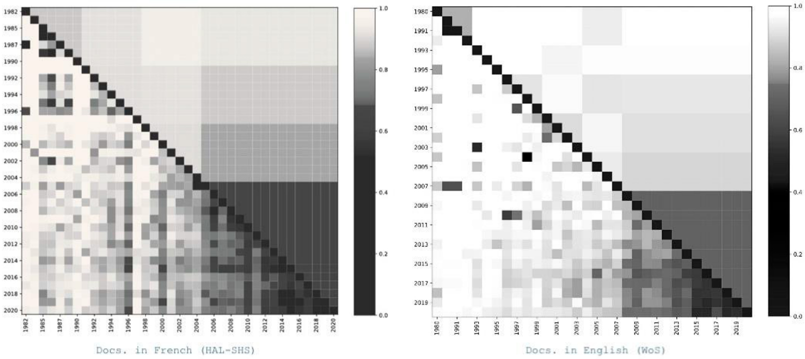


Figure 2. Period detection and significant term correlation between years in French and English.

The underlying semantic structure, summarized in Figure 2, allows us to understand how this growth unfolds. It charts how many terms used in one year resemble the one from the other years in the series. The whiter the cell is, the most dissimilar those two years are. On each graphic's upper right, detected most similar periods are represented by using a homogeneous grayscale. In French, three short periods, between four and eight years of duration, are detected. The fifth and last is the one that covers more time, from 2005 to 2020. It is also the one with documents that resemble the most to each other. In English, the similarity between the papers over the years is more irregular and weak. Here, detected periods are more -six in total- and also shorter. In general, the debate in English is more loose and erratic.

Discussion

Results have shown how similar topics can be addressed at different paces in two different languages. As the volume of documents rises first in French and, later, in English, it is natural to assume that many of the ideas developed in one language get translated and adapted to the other. Our results on the semantic structure of the debate in each language confirm this ‘'natural'' pattern: the production in French has a tighter, firmer, and stronger underlying structure. Topics appear on a longer time frame, a fact we could relate to long prolonged and sustained tradition on these subjects with a steady amount of documents published in the first of the three periods. The last detected period of fifteen years (2005-2020) speaks of a stabilization of the researchers' main concerns interests reflected in Ffrench documents. The delay between French and English-language publications on the same topics relates can be explained by a valorization strategy done by publishing English-language articles or books only after a valorization strategy that takes place after a maturation period that is visible in the French-speaking literature. Matters of interest get first widely discussed in French and then published in English.

We could also use the metaphor of ‘carry publishing’ (as we talk of carry trading in the financial investment business) since we can consider the articles in French as a local investment maturing over time to benefit in the long term by a publication in English on the international arena. It has potential implications for a better understanding of knowledge circulation dynamics and language interaction between English and non-English speaking scientific literatures (Kulczycki et al., 2018). This process takes place over the long term (Gordin, 2015) and relates to both scientific strategic choices and structural social constraints in the production of scientific knowledge (Hanafi and Arvanitis, 2014). The flat slope in publication rate from 2018 and on allows us to assume that these ‘carry publishing’ strategies have limits. Here, novelty appears as an inherent solid constraint since the international interest may decrease over time.

Further research should investigate these publication practices for non-English speakers; the particular case of French is of specific interest since it goes well beyond France (e.g., French-speaking Africa and North-African countries publish more in French than French scientists). How much does the thematic orientation of French documents resemble that of the English ones? What role authors’ choices play in this process? How do research institutions, evaluation standards, and international competition (as reflected in universities’ rankings) impact the publication process?

Conclusions

This research has shown that, on issues of global relevance, language matters. Specific interests can develop in a particular way in specific linguistic and geographic boundaries. These perspectives do eventually migrate to the English language but in a less systematic way. Beyond observing the fact that social sciences publish more heavily in their “national” languages than other scientific domains, little research has been done on the fundamental issue of language (Gordin, 2015; Ortiz, 2008; Kulczycki et al., 2018. 2020).

Our results indicate a maturation process over time, which affects the value given to locally-published scientific research since the English language is predominant in the core-publishing journals (Hanafi, 2011; Keim, 2016). Regarding the method used, a research strategy based on keywords complemented with a dialogue with researchers reveals a strong potential for a strategic-oriented analysis. Since informants may not be fully aware of these language-processing techniques, they might have difficulties providing precise feedback when examining a resulting semantic map. A strong interaction along the analytical process is needed to obtain fine-grained results and build query lines in the data-mining process. Last, the local perspective in our analysis allowed us to link the scientific debates at the national level with those appearing in English journals. It has been possible in the case of France, where local data were available, and because French social sciences also circulate beyond the national boundaries (see, e.g., 'French theory' in the USA). But as national repositories and databases are growing everywhere (see the databases such as Latindex in Latin America, or the new repositories in Asia, China, Russia, India), studying differences between locally published research in non-English speaking contexts and English-speaking international authors will be feasible and will probably reveal specific determinants that go beyond the need to diffuse more widely research results.

Acknowledgments

Research funded by the IdEx platform, from the Université de Paris, and the GRIP.

References

Arvanitis, Rigas et David O'Brien, Eds. (2019). *The Transformation of Research in the South: policies and outcomes*. Paris: Archives Contemporaines & IRD.

CCSD. (2021). API HAL. API Archive Ouverte HAL [Respository]. Aarchive ouverte HAL-SHS (Sciences de l'Homme et de la Société). https://api.archives-ouvertes.fr/docs/search

Chi, P.-S. (2012). Bibliometric characteristics of political science research in Germany. *Proceedings of the American Society for Information Science and Technology*, 49(1), 1–6.

Cruz-Castro, L., & Sanz-Menéndez, L. (2018). Autonomy and Authority in Public Research Organisations: Structure and Funding Factors. *Minerva*, 1–26.

Engels, T. C. E., Ossenblok, T. L. B., & Spruyt, E. H. J. (2012). Changing publication patterns in the Social Sciences and Humanities, 2000–2009. *Scientometrics*, 93(2), 373–390.

Gordin, Michael D. (2015). *Scientific Babel. How Science Was Done Before and After Global English*: U of Chicago Press.

Hanafi, S. & Arvanitis, R. (2014). "The marginalization of the Arab language in social science: Structural constraints and dependency by choice." *Current Sociology* 62(5): 723-742.

Keim, W. (2016). "The international circulation of social science knowledge. Relevant factors for acceptance and rejection of travelling texts." Revue d'anthropologie des connaissances 10(1): a-aj.

Kulczycki E, Guns R, Pölönen J, et al. (2020) Multilingual publishing in the social sciences and humanities: A seven-country European study. J. Assoc. Inf. Sci. Technol. 71(11), 1371–1385.

Kulczycki, E., Engels, T.C.E., Pölönen, J. *et al.* (2018) Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics* 116, 463–486.

Ortiz, R. (2008). *La supremacía del inglés en la ciencias sociales*.México: Siglo XXI. 236 p.

Ràfols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, 61(9), 1871–1887. https://doi.org/10.1002/asi.21368

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411–423.

Wallace, M., & Ràfols, I. (2015). Research portfolio analysis in science policy: Moving from financial returns to societal benefits. *Minerva*, 53(2), 89–115.

Wallace, M., & Ràfols, I. (2018). *Institutional Shaping of Research Priorities: A Case Study on Avian Influenza* (SSRN Scholarly Paper ID 2745137). Social Science Research Network.

Zitt, M., & Bassecoulard, E. (1994). Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis. *Scientometrics*, 30(1), 333–351.

[i] Operations were conducted to clean terms. Void terms were excluded. Specific terms were homologated manually when the software had not detected their sinonimity. Foreign terms were kept when related to objects of interest. 257 terms were deleted or homologated of the French collection. In English, a total of 675 terms remained.

RRI projects in the mirror of RRI indicators

Ülle Must[1]

[1] *ulle.must@harno.ee*
Estonian Education and Youth Board, Ülikooli 2, 51003 Tartu (Estonia)

Abstract

The Responsible Research and Innovation (RRI) approach encourages societal actors to work together during the whole research and innovation (R&I) process to better align R&I and its outcomes with the values, needs and expectations of the society. In this paper, we monitor the results of the implementation of the concept of responsible research and innovation, primarily from the perspective of the world's largest funding programme – the EU Framework Programme RRI related projects. The analysis is primarily based on data on the funded projects related to RRI under the EU Framework Programmes FP7 and Horizon 2020, in total 85 projects with 845 participants (organisations). For the sake of completeness of the analysis (unified database structure), we used the Clarivate Web of Science database. In total we retrieved 212 papers. We used RRI indicators developed in the MoRRI project for testing project activities. In the context of this paper, we take a closer look at the implementation of the four principles: public engagement, gender balance, open access, and governance.

Keywords
RRI, EU Framework Programme, MoRRI indicators

Introduction

The notion that "soft" values have a role to play at all stages of society's life only took root in the last quarter of the 20th century, and it did not receive its "price tag" until the second decade of the 21st century.

The term “responsible research = RR”, or “responsible innovation = RI”, or “responsible research and innovation = RRI” is not new, and in publications dates back from the 80s of the 20th century (Figure 1).

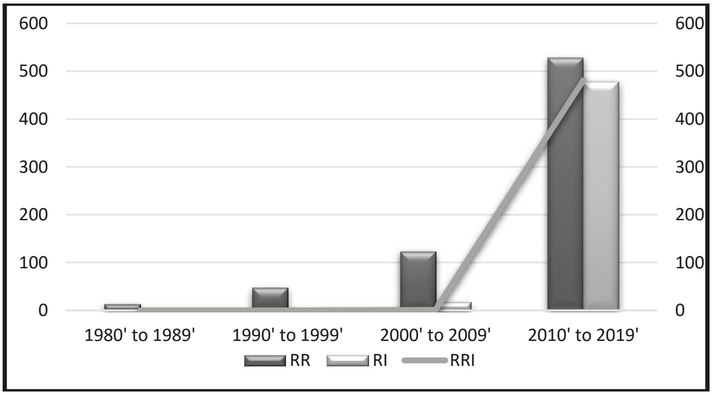


Figure 1. RR, RI, and RRI papers by timeline (WoS as of 01.01.2021)

The concept of RRI gained real impetus after the European Commission started to support it in 2011 (Sutcliffe, 2011, von Schomberg, 2011). As stated by Tim Flink and David Kaldewey (Flink and Kaldewey, 2018), “STI policy discourses in Europe and elsewhere have seen the arrival and uptake of new concepts, models and metaphors, all of which seem to challenge the old idea of an overarching social contract between science and society as well as those socioeconomic narratives of planning that surrounded the linear model of innovation”. This is well illustrated by the linguistic analysis of the basic documents of the EU Framework Programme (FP), which show that allusions to "soft" values did not appear in FP official texts until the seventh year of its operation, since FP3 (1990-1994), and gained real substantive value from FP7 (2007-2014) (Mustajoki, et al, 2020).

RRI may be defined as an ‘on-going process of aligning research and innovation to the values, needs and expectations of society.’

The RRI approach encourages societal actors to work together during the whole research and innovation (R&I) process to better align R&I and its outcomes with the values, needs and expectations of the society. It is like an umbrella that brings together different aspects of collaboration, ensuring the participation of different stakeholders in research and development (R&D), open access to the process and results, gender equality, timeliness in science education, ethics and good governance. As the ambition was to bring together under common denominator practitioners, policy makers, researchers, entrepreneurs and the so-called “aunt Maali” from the countryside, there were difficulties in making the term understandable. Because of that several authors expressed their concern on the risk that RRI will remain a vague set of hopeful 'Keys' which must be incorporated into funding proposals but do not significantly influence the norms, discourses and functions of other institutions in the EU, including those involved in regulating the end products of scientific research (de Saille, 2015). An opinion was also issued that the Horizon Swafs/RRI expert advisory group is lacking a clear idea of what RRI ‘is’ (Rip, 2016). A position was also presented that the main obstacle for RRI integration appears to be the policy integration strategy itself: the RRI framework is not clear to those who are the intended users which hinders the effective operationalization of RRI in research practice (Braun, et al, 2019).

Over time, it has been recognized that RRI will inevitably mean different things to different people, and demand different forms of engagement in different countries, cultures and scientific disciplines (Peter, et al, 2018), and the focusing on the term “RRI” creates unnecessary barriers, so, it is better to concentrate on the meaning and principles of RRI (Delaney, et al, 2020).

In this paper, we monitor the results of the implementation of the concept of responsible research and innovation, primarily from the perspective of the world's largest funding programme – the EU Framework Programme RRI related projects (Table 1). We used RRI indicators developed in the MoRRI project (www.morri-project.eu) to monitor the implementation of the four RRI key elements (gender, engagement, open access, governance) in RRI projects themselves.

Table 1. FP7 and H2020 RRI calls (Horizon Dashboard as of 01.01.2021)

Topic	*Topic*	*Signed Grants*
FP7-Adhoc-2007-13		2
FP7-SCIENCE-IN-SOCIETY-2012.1.1.1-1	Governance frameworks for Responsible Research and Innovation (RRI)	2
FP7-SCIENCE-IN-SOCIETY-2012.1.2.1-1	International Coordination in the field of Responsible Research and Innovation (RRI)	4

FP7-SCIENCE-IN-SOCIETY-2012.1.3.3-1	Scientific data: open access, dissemination, preservation and use	1
FP7-SCIENCE-IN-SOCIETY-2013.1.1.1-1	Production and use of a Training and Dissemination Toolkit on Responsible research and innovation	1
FP7-SCIENCE-IN-SOCIETY-2013.1.1.1-2	Responsible Research and Innovation in industrial context	1
FP7-SCIENCE-IN-SOCIETY-2013.2.2.1-1	Raising youth awareness to Responsible Research and Innovation through Inquiry Based Science Education	4
H2020-FETOPEN-2-2014	Coordination and Support Activities	1
H2020-GARRI-1-2014	Fostering RRI uptake in current research and innovations systems	2
H2020-GARRI-2-2015	Responsible Research and Innovation in industrial context	3
H2020-ICT-35-2016	Enabling responsible ICT-related research and innovation	5
H2020-ISSI-1-2014	Pan-European public outreach: exhibitions and science cafés engaging citizens in science	1
H2020-ISSI-1-2015	Pan-European public outreach: exhibitions and science cafés engaging citizens in science	2
H2020-ISSI-2-2014	Citizens and multi-actor engagement for scenario building	1
H2020-ISSI-3-2015	Knowledge Sharing Platform	1
H2020-ISSI-5-2014	Supporting structural change in research organisations to promote Responsible Research and Innovation	1
H2020-ISSI-5-2015	Supporting structural change in research organisations to promote Responsible Research and Innovation	3
H2020-NMP-32-2015	Societal engagement on responsible nanotechnology	1
H2020-SEAC-1-2014	Innovative ways to make science education and scientific careers attractive to young people	2
H2020-SEAC-2-2014	Responsible Research and Innovation in Higher Education Curricula	2
H2020-SwafS-01-2016	Participatory research and innovation via Science Shops	2
H2020-SwafS-04-2016	Opening Research Organisations in the European Research Area	2
H2020-SwafS-05-2017	New constellations of Changing Institutions and Actors	2
H2020-SwafS-05-2018-2019	Grounding RRI practices in research and innovation funding and performing organisations	4
H2020-SwafS-06-2017	Engaging industry – Champions for RRI in Industrial Sectors	1
H2020-SwafS-09-2016	Moving from constraints to openings, from red lines to new frames in Horizon 2020	1
H2020-SwafS-12-2017	Webs of Innovation Value Chains and Openings for RRI	1
H2020-SwafS-13-2017	Integrating Society in Science and Innovation – An approach to co-creation	2
H2020-SwafS-14-2017	A Linked-up Global World of RRI	1
H2020-SwafS-14-2018-2019-2020	Supporting the development of territorial Responsible Research and Innovation	11
H2020-SwafS-15-2018-2019	Exploring and supporting citizen science	7
H2020-SwafS-17-2019	Consolidating and expanding the knowledge base on citizen science	1
H2020-SwafS-18-2018	Taking stock of the application of the precautionary principle in R&I	1
H2020-SwafS-20-2018-2019	Building the SwafS knowledge base	4

H2020-SwafS-21-2018	Advancing the Monitoring of the Evolution and Benefits of Responsible Research and Innovation	1
H2020-SwafS-23-2017	Responsible Research and Innovation (RRI) in support of sustainability and governance, taking account of the international context	1
H2020-SwafS-23-2020	Grounding RRI in society with a focus on citizen science	1
H2020-SwafS-27-2020	Hands-on citizen science and frugal innovation	1
H2020-SwafS-31-2020	Bottom-up approach to build SwafS knowledge base	3

As we can see in Table 1, the range of issues intended to be addressed in RRI projects is wide: implementing institutional changes in different types of organisations (research funding and performing organisations, higher education institutions, etc.), based on RRI principles; creation of practical tools and best practice examples that directly address societal needs while contributing to environmental and economic sustainability in industry-focused projects; the projects supporting EU regions to develop more open and collaborative approaches to society by taking an RRI approach; plus projects with disciplinary or sectoral approaches (e.g. focused on marine research institutes, the biosciences, energy sector or deindustrializing regions (Delaney, et al, 2020).

Methodology

Data sources

The analysis is primarily based on data on the funded projects related to RRI under the EU Framework Programmes FP7 and Horizon 2020, in total 85 projects with 845 participants (organisations). RRI as a concept in itself does not feature in the proposal template or in the evaluation sub-criteria. If one of the dimensions of RRI (public engagement, gender, ethics, open access, science education) is pertinent, then the topic is flagged for RRI in Horizon Dashboard (Delaney, 2020). One way to monitor the performance of projects is to analyze the publications. As these are projects that require the involvement and synergy of all parties, certain limitations must be taken into account when interpreting the results. From H2020, bibliographic data on project publications will also be publicly available[1], where we see that the number of published publications during Horizon 2020 period is significant (350,000 papers in total, among them 341,000 peer-reviewed papers). On closer inspection, we see that the lion's share of the publications have been completed by the recipients of the ERC grant (198,400 in total, from the 147,600 peer-reviewed). The projects funded by the Science with and for Society (which mainly funded the RRI projects – see Table 1) published a total of 350 publications, of which 341 were peer-reviewed (Horizon Dashboard). For the sake of completeness of the analysis (unified database structure), we used the Clarivate Web of Science database. To perform the search, we used a combination of the WoS fields "funding organisation", "Funding Text ", "Grant number", "Author keyword" and "Keywords Plus". In the interest of data accuracy, a manual check of the data was finally carried out. In total we retrieved 212 papers. The discrepancy of the obtained data with the data in the Dashboard can be explained either by the lack of a correct form of acknowledgments of the author(s) (according to the Horizon 2020 Annotated Model Grant Agreement, Article 38.1.2) or by the fact that the journal title is not covered in the Master Journal List of Clarivate WoS.

[1] In April 2021 the European Commission's open access publishing platform *Open Research Europe (ORE)* was officially launched

Analysis

For data analysis, we used combined data from two datasets. We used the following fields for project data: a) projects: start and end date, funding scheme; b) organisations: name, country, activity type and the role in the project. From publication data we used the following fields: Author Full Names; Author Keywords; Keywords Plus; Funding Organisation; Publication Year; Open Access Designations.

MoRRI indicators

We used selection of RRI indicators proposed by MORRI project (*Monitoring the Evolution and Benefits of Responsible Research and Innovation*) funded by the EU Framework Programme. The project's main objective was to provide scientific evidence, data, analysis and policy intelligence to support directly Directorate General for Research and Innovation (DG-RTD) research funding activities and policy-making activities in relation with Responsible Research and Innovation (RRI). At the end of the project, 36 RRI indicators were proposed for use. In the context of this paper, we take a closer look at the implementation and compliance of the four RRI principles in the projects.

Findings

General findings

A characteristic feature of FPs is their long-term duration, in terms of both signed contracts and the output timeline. Based on the data we can say that in FPs immediate results should not be expected. As in some areas the impact of the project has lasted for decades (Must, 2017).

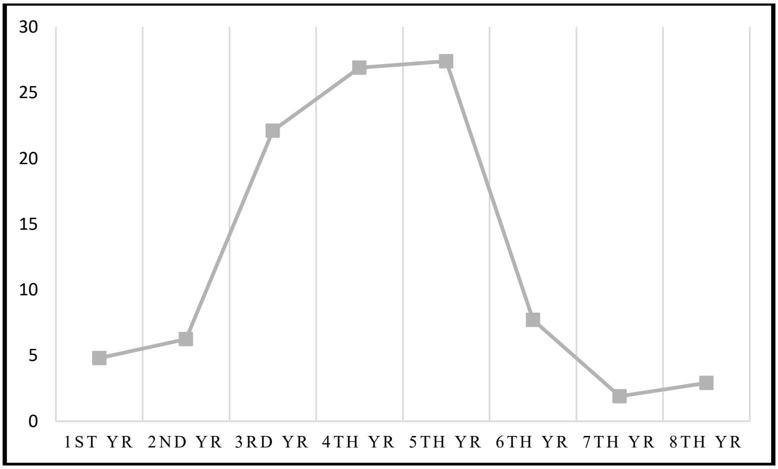


Figure 3. The Frequency of publications of RRI projects during the project time cycle (%) (WoS and Horizon Dashboard as of 01.01.2021)

The continued validity of this statement was also shown by the data of this analysis – the lion's share of project papers have been published in the fourth and fifth year of the project (54.3%) when the majority of projects have already been completed.

Most of the funded projects (63.5%) are funded under the funding scheme *Coordination and support action* (= CSO) which means that it is expected that during the project such activities are conducted as dissemination, awareness-raising and communication, networking, coordination or support services, policy dialogues and mutual learning exercises, also minor studies. In *Research and Innovation actions* (RIA) (34.1% of funded projects) the activities foreseen aim to establish new knowledge and/or to explore the feasibility of a new or improved technology, product, process, service or solution. For this purpose they may include basic and applied research, technology development and integration, testing and validation on a small-scale prototype in a laboratory or simulated environment (Horizon 2020 Online Manual). Against this background, it is somewhat surprising that the data from the present study show that CSA project partners outperformed RIAs in terms of publishing activity.

Public engagement

Traditionally, research and education institutions have been the most active participants in the Framework Programme. From the point of view of RRI aims, the extent to which representatives of business, policy makers or civil society have been involved, is crucial. As we see from Figure 4, representatives of higher or secondary education establishments (HES), research organisations (REC), and other bodies (OTH) constitute the majority of partners in the RRI related projects. At the same time, if we compare these data with overall data, we see some tiny shift to greater involvement of private for-profit entities (PRC) and public bodies (PUB). In terms of EU FP terminology, the organisation type "Other" forms a very heterogeneous community – this includes institutions (foundations) established by the state, political parties, stakeholders, societies. RRI projects have significantly more representatives of these organisations than the average of all projects. As expected, more experienced partners belong to universities, one fifth of HES organisations are involved in more than one RRI projects, in total 23.8% of organisations participate in more than one RRI project. This shows that the core of so-called RRI researchers has developed who involve a wider range of new partners in their projects.

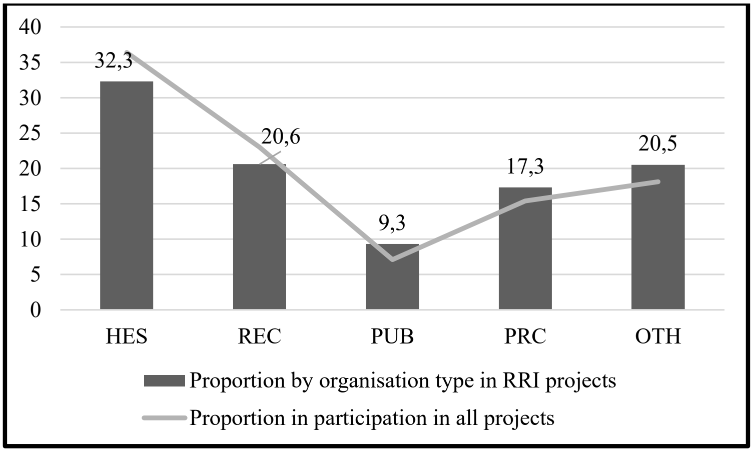


Figure 4. FP project partners by organisation type (Horizon Dashboard as of 01.01.2021)

About ten years of RRI funding allows us to highlight partnerships that have lasted in more than one project: *Applied Research and Communications Fund* (BG) / *Danish Board of Technology Foundation* (DK) have cooperated in five common projects. In four common projects *Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.v.* (DE) / *Institute for Advanced Studies* (AT) and *Zentrum für Soziale Innovation GmbH* (AT) / *Universiteit Leiden* (NL). Three projects link five organisations: *Aarhus Universitet* (DK) / *Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.v.* (DE) / *Institute for Advanced Studies* (AT) / *Zentrum für Soziale Innovation GmbH* (AT) / *Conoscenza e Innovazione Societa A Responsabilita Limitata Semplificata* (IT). There is no doubt that established partnerships have a positive side but one of the ideas of FP funding is its multiplier effect – the involvement of different new partners makes it possible to consolidate ideas and thus ensure the sustainability of the field. As stated in MoRRI, "There is still a long way to go regarding the 'universe' of researchers in Europe before RRI is more broadly known and accepted." Researchers receiving funding from the EU framework programme are more familiar with the concept of RRI, and they also associate more future benefits than non-funded ones (Mejlgaard, et al, 2018). Operating in a closed circle makes it possible to claim that the feedback from project to policy-making is arguably weak (Nedozhogina and Hõrak, 2019). At the same time we have to take into account that processes (on both sides) take time in RTD. It also means a global introduction to the concept of RRI. As we see from Table 2, the lion's share of project participants and authors to RRI papers come from Europe. However, as we already stated previously, different countries may not call similar activities with the same name, so for example a similar concept of RRI is called "Scientific social responsibility" in India (Braun, et al, 2020), and the concept of RRI is more entrenched in Europe than the rest of the world.

It means that there is a need to adapt existing frameworks so that they can operate within a country's technological, social and political context, while at the same time emphasizing the international environment and the global governance of innovations (Gao, and Zhao, 2019).

Table 2. Participants and authors by geographic area (%)
(WoS and Horizon Dashboard as of 01.01.2021)

Geographic area	*Project Participants*	*Authors*
Africa	0.5	1.1
Asia	1.2	3.8
Australia-Oceania	0.1	2.4
Europe	96.7	87.1
N-America	0.4	3.8
n/a	0	0.5
S-America	1.1	1.3

Gender equality

Gender equality and diversity is not just a matter of social justice, it leads to higher quality science, and is also a component of "smart economics". It is also a well-known fact that gender-mixed teams perform better. In H2020 several indicators are used for the monitoring of gender equality – participation in projects, being project coordinators or members of high level EC advisory groups, etc. On an average, women represent 42% of the participants in projects (including non-researchers) and 28% of projects have women coordinators (Gender equality and diversity, 2019). From this point of view, RRI projects far exceed the results of the FP – half of the project coordinators are women (Figure 5).

The monitoring of research outputs by gender is one indicator which is often used by research funding organisations. The number and proportion of female authors highlights developments in women's representation across fields and sectors over time, on the basis of bibliometric data. The MORRI data show that in 2016 the average proportion of female authors was below 40% (Mejlgaard, et al, 2018). As to papers written by RRI project participants, the gender of authors is almost in balance – 49.5% are female authors, 50.5% are male authors. However, it should be mentioned that women are slightly more productive than men – the average production of women is 1.5 papers, and that of men 1.4. At the same time, the corresponding authors' data are strongly inclined towards male authors (57% vs. 43%) (Figure 5).

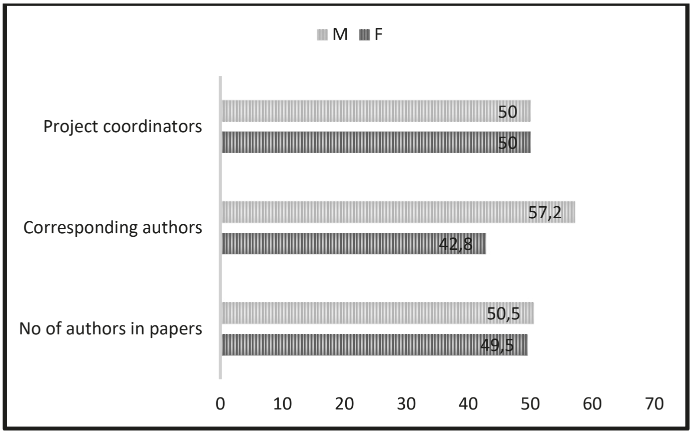


Figure 5. The distribution of RRI papers' authors, corresponding authors and RRI project coordinators by gender (%) (WoS as of 01.01.2021)

When dealing with gender-specific data, a number of factors must be taken into account that may influence results, such as the authors' field of research, working experience, etc.

Open access

Open access is the idea of making research results freely available to anyone who wants to access and re-use them, and thus is one of the most accepted keys of RRI. Originally, open access was separated into 'gold' and 'green' where gold indicates open access journals and green indicates open access through self-archiving. In current survey we use Clarivate WoS data where papers are labelled with three different types of open access (Clarivate):

- Gold: a) DOAJ Gold – Articles published in journals listed in the Directory of Open Access Journals (DOAJ). To be listed in the DOAJ, all articles in these journals must have a license in accordance with the Budapest Open Access Initiative; b) other gold – articles have a Creative Commons (CC) license by Impactstory's Unpaywall Database but are not in journals listed in the DOAJ. Most of these articles are from hybrid journals. Hybrid open access journals are subscription journals that include some open access articles.
- Bronze: the licensing for these articles is either unclear or identified by Impactstory's Unpaywall Database as non-CC license articles. These are free-to-read or Public Access articles located on a publisher's site.

- Green: a) Published – final published versions of articles hosted in an institutional or subject-based repository, b) Accepted – Accepted manuscripts hosted in a repository. Content is peer reviewed and final but may not have undergone the publisher's copyediting or typesetting.

The results of this study show that 68% of RRI articles are OA publications. The most common combination is DOAJ Gold and Green Accepted (33.6% of cases), followed by Green Published and Other Gold (19.6%).

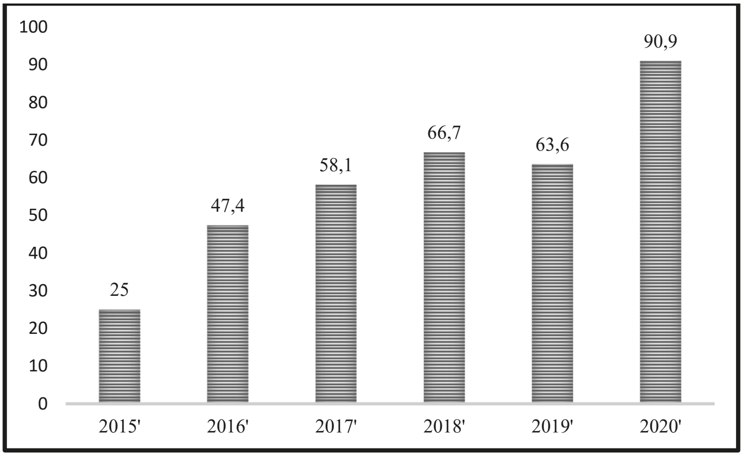


Figure 6. The proportion of OA RRI papers by year (%) (WoS as of 01.01.2021)

As we see from Figure 6, there is a constant growth trend in the distribution of OA publications. It was expected also in MoRRI, when analysing the data for 2012-2015 that a general trend will increase over time. The average annual EU-28 growth rate of the shares for 2012-2015 was 26% (Mejlgaard, et al, 2018).

Governance

The governance dimension fosters institutional transformations, developing conducive framework conditions for RRI, and supporting changing cultures and practices of research and innovation actors. It is defined as a way in which societal and state actors intentionally interact in order to transform ST&I systems, by regulating issues of societal concern, defining processes and direction of how technological artefacts and innovations are produced, and shaping how these are introduced, absorbed, diffused and used within society and economy. In context of RRI, governance mechanisms examine whether research-funding (RFO) and performing organisations (RPO) have established processes for managing the key areas of RRI.
For the present study, we used WoS data of RRI projects to monitor RRI acceptance by funding organisations. In total 294 funding organisations were acknowledged in the papers, from them 86 unique funders. The results showed that more than a half (55%) of the papers were written solely with the support of the EU Framework Programme. In addition to FP support, other EU-based funding sources (COST, ERDF, INTERREG, and Structural Funds) were used. A total of 60% of the articles were published thanks to various EU funding sources.

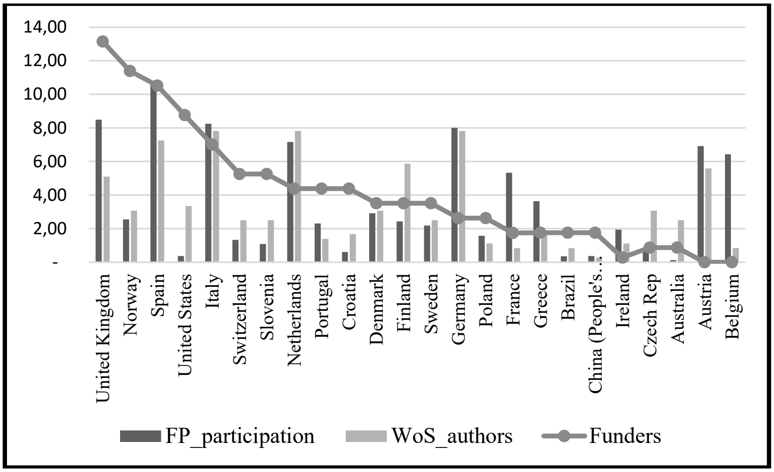


Figure 7. The proportion of FP participation, authors of the papers and acknowledged funders by country (WoS and Horizon Dashboard as of 01.01.2021)

Participants in RRI projects were distributed among 54 countries, of which authors from 38 countries were on the list of authors, and national funding organisations from 22 countries were mentioned in the acknowledgements.

Comparing the proportions of participants in projects, authors of papers and funders, we see certain patterns (Figure 7). The proportion of participation in projects does not necessarily mean a higher proportion of participation among the authors of the papers. This is most proportional for Spain and Germany. However, there are proportionally more papers compared to participating in projects in the Netherlands, and Finland. The list of funding organisations offers interesting results – in addition to the so-called special suspects (UK, NO, ES, US, IT), Slovenia and Croatia are among the top seven funders. At the same time, Austria and Belgium who belong to the top seven project participants, carry out their RRI projects only on the basis of EU Framework Programme funding. Among the list of the five most mentioned funders are the Norwegian Research Council (e.g. Programme on Responsible Innovation and Corporate Social Responsibility), U.S. National Science Foundation, Slovenian Research Agency, National Institute for Health (UK), and Croatian Science Foundation.

Despite the fact that the vast majority of research and innovation (R&I) is funded and produced by industry, companies tend to have no awareness or recognition of this concept (Gurzawska, et al, 2017), and in case of the current survey, only one company was acknowledged among a list of funders.

Discussion

In the present study, we tested what RRI projects look like against the background of MoRRI indicators. Public engagement is the most problematic, as research and education institutions dominate among the partners, and the involvement of other organisations in consortia is quite marginal. At the same time, comparing the data in the general environment, we observe a significantly higher participation of business and government organisations in projects than in the Framework Programme on average.

With regard to gender equality, we can see even more positive trends: the gender of RRI papers' authors is almost in balance – 49.5% are female authors, 50.5% are male authors. When dealing with gender-specific data, a number of factors must be taken into account that may influence results, such as the authors' field of research, working experience, etc.
Open access publication is a growing trend accounting for an average of 68% of all papers. Taking into account that governance mechanisms examine whether research-funding (RFO) have established processes for managing the key areas of RRI, our data show that EU funding plays a key role, and the impact of national funding is low.
The ten years of RRI funding by the European Commission have undoubtedly given a boost to this direction. Horizon Europe calls results will show how sustainable the established RRI partnerships are in the context when the RRI is integrated into the overall context.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 741402.

References

Braun, R., Gianni, R., Srinivas, K. R. (2020). Policy transfer and shared knowledge base learning from policy implementation. Responsible research and innovation (RRI) in the European Union and Scientific social responsibility (SSR) in India. *NewHorrizon Policy Brief*, 3, 1-5.

Braun, R., Loeber, A., Novitzky. P. (2019). Lacking integration of societal needs and ethical concerns into European research and innovation policy severely limits the ability to tackle 'grand challenges'. *NewHorrizon Policy Brief,* 2, 1-4.

Clarivate. Retrieved on 01.01.2021 from http://info.clarivate.com/openaccess

Delaney, N., Tornasi, Z., Iagher, R., Monachello, R., Warin, C. (2020). Science with and for Society in Horizon 2020 Achievements and Recommendations for Horizon Europe. Luxembourg: Publications Office of the European Union. 129 pp.

de Saille, S. (2015). Innovating innovation policy: the emergence of 'Responsible Research and Innovation'. *Journal of Responsible Innovation*, 2 (2). 152 – 168.

Flink, T., Kaldewey, D. (2018). The new production of legitimacy: STI policy discourses beyond the contract Metaphor. *Research Policy*, 47 (1), 14-22.

Gao, L; Liao, M; Zhao, YD. (2019). Exploring complexity, variety and the necessity of RRI in a developing country: the case of China. *Journal of Responsible Innovation*, 6 (3), 368-374.

Gender equality and diversity in R&I. (2019). Retrieved on 01.01.2021 from https://ec.europa.eu/info/sites/info/files/research_and_innovation/knowledge_publications_tools_and_data/documents/ec_rtd_factsheet-gender-equality_2019.pdf

Gerber, A., Forsberg, E.M., Shelley-Egan, C., Arias, R., Daimer, S., Dalton, G., Cristbal, A B., Dreyer, M., Griessler, E., Lindner, R., Revuelta, G., Riccio, A., Steinhaus, N. (2020). Joint declaration on mainstreaming RRI across Horizon Europe. Journal of Responsible Innovation, 7 (3), 708-711.

Gurzawska, A; Makinen, M; Brey, P. (2017). Implementation of Responsible Research and Innovation (RRI) Practices in Industry: Providing the Right Incentives. *Sustainability*, *9*(10), 1759, https://doi.org/10.3390/su9101759

Horizon 2020 Online Manual. Retrieved on 01.01.2021 from https://ec.europa.eu/research/participants/docs/h2020-funding-guide/grants/applying-for-funding/find-a-call/what-you-need-to-know_en.htm

Horizon Dashboard. Retrieved on 01.01.2021 from https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/horizon-dashboard

Mejlgaard, N., Bloch, C., B. Madsen E. B., Griessler, E., Wuketich, M., Meijer, I., Woolley, R., Lindner, R., Bührer, S., Jäger, A., Tsipouri, L., Stilgoe, J. (2018). Monitoring the evolution and benefits of responsible research and innovation in Europe. Summarising insights from the MoRRI project. Luxembourg: Publications Office of the European Union. 74 pp.

Must, Ü. (2017). Bibliometric Study on the sustainability of EU FP projects: Methodological aspects. *Collnet Journal of Scientometrics and Information Management*, 11 (2), 215-222.

Mustajoki, A., Mustajoki, H., Must, Ü. (2020). Vastuullista tiedettä. *Tieteessä tapahtuu*, 1, 3-12.

Nedozhogina, O., Hõrak, H. (2019). RRI implementation in Horizon 2020 and the future of RRI in Horizon Europe. *Hubit Policy Brief*, 4. 1- 9.

Peter, V., Mejlgaard, N., Bloch, C., Madsen, E. B., Griessler, E., Wutekich, M., Meijer, I., Wooley, R., Lindner, R., Bührer, S., Jäger, A., Tsipouri, L., & Stilgoe, J. (2018). Monitoring the evolution and benefits of Responsible Research and Innovation in Europe: Summarising insights from the MoRRI project. Brussels: European Commission. 71 pp.

Rip, A. (2016). The clothes of the emperor. An essay on RRI in and around Brussels. *Journal of Responsible Innovation*, 3 (3), 290-304.

She figures 2018. (2019). Luxembourg: Publications Office of the European Union. 216 pp.

Sutcliffe, H. (2011). A report on Responsible Research & Innovation. 34 pp. Retrieved on 01.01.2021 from http://www.diss.unimi.it/extfiles/unimidire/243201/attachment/a-report-on-responsible-research-innovation.pdf

von Schomberg, R. (2011). Research and Innovation in the Information and Communication Technologies and Security Technologies Fields: A Report from the European Commission Services. (R. von Schomberg, Ed.). European Union, Publications Office of the European Union, Luxembourg.

Diversity and interdisciplinarity - Should variety, balance and disparity be combined as a product or better as a sum? A probability-theoretical approach

Rüdiger Mutz

ruediger.mutz@uzh.ch
Center of Higher Education and Science Studies, University of Zurich,
Andreasstrasse 15, 8050 Zurich, Switzerland

Abstract

Diversity is a central concept not only in ecology, but also in the social sciences and in bibliometrics. The discussion about an adequate measure of diversity is strongly driven by the work of Rao (1982) and Stirling (2007). It is to the credit of Leydesdorff (2018) to have proposed a decisive improvement with regard to an inconsistency in the Rao-Sterling-diversity indicator that Rousseau (2018) had pointed out. With recourse to Shannon's probabilistically based entropy concept, in this contribution the three components of diversity "variety", "balance", and "disparity" are to be reconceptualized as entropy masses that add up to an overall diversity indicator div_e. Diversity can thus be interpreted as the degree of uncertainty or unpredictability. For "disparity", for example, the concept of *mutual information* is used. This overall probability-theoretical based concept is applied exemplarily to data on research output types of funded research projects in UK that were the subject of the Metric Tide Report (REF 2014). As expected, research output types depend on the research area, with journal articles having the strongest individual balance among the output types, i.e. being represented in almost all research areas. However, the overall diversity is comparably high (78.8% of the maximum diversity).

Introduction

Diversity is not only a central concept in scientometrics (e.g., interdisciplinarity), but also in ecology (e.g., Jost, 2007) or in the social sciences (e.g., Brusco, Cradit, & Steinley, 2020; Jones & Dovidio, 2018. Therefore, the measurement and operationalization of this concept is of eminent importance, which is also reflected in numerous publications on this concept especially in the field of ecology. A certain break in the development of this concept comes from the work of Rao and Stirling (Rao, 1982; Stirling, 2007), who developed a diversity concept based on the existing literature, which, in brief, consists of three components:

- *Variety*: «Variety is the number of categories into which system elements apportioned. It is the answer to the question: "How many types of things we have? … All else being equal, the greater the variety, the greater the diversity» (Stirling, 2007, p. 709)
- *Balance*: «Balance is a function of the pattern of apportionment of elements across categories… All else being equal the more even is the balance, the greater the diversity.» (Stirling, 2007, p. 709)
- *Disparity*: "Disparity refers to the manner and degree in which the elements may be distinguished… All else being equal, the more disparate are the represented elements, the greater the diversity" (Stirling, 2007, p. 709).

Thus, first Rao (1982) and later Stirling (2007, p. 721) developed a "general diversity heuristic" as a sum indicator of different elements j, i of a system (e.g., different research area in research proposals):

$$D = \sum_{ji(i \neq j)} d_{ij} p_i p_j \ . \tag{1}$$

where p_i and p_j are proportions of elements i and j in the system (as base for balance), and d_{ij} is the is the degree of difference between the elements (disparity). Variety is the number of different elements. This concept had a strong influence not only in ecology (e.g., Rousseau, Van Hecke, Nijssen, & Bogaert, 1999) but also in scientometrics (e.g., Goyanes, Demeter, Grané, Albarrán-Lozano, & Gil de Zúñiga, 2020; Rousseau, 2018; Wang, Thijs, & Glänzel, 2015) The idea of composing an indicator multiplicatively from several individual indicators is captivating and opens up many possibilities for analysis. It is also easy to calculate.

Unfortunately, Rousseau (2018, p. 651) was able to show with a simple data example that a central assumption of the diversity indicator ("monotonicity") formulated by Stirling (2007, p. 711) does not hold. For a given variety and disparity the measure does not increase monotonically with balance. This points to fundamental problems with this concept. It is to the credit of Leydesdorff (2018) and Leydesdorff, Wagner, and Bornmann (2019b) to create an indicator that does not have these problems and continues to multiplicatively link the three individual diversity indicators, which is certainly a viable approach despite criticism and modifications (Leydesdorff, Wagner, & Bornmann, 2019a; Rousseau, 2019).

For i=1 to n_c and j=1 to n_c categories, the diversity is defined as follows (Leydesdorff, 2018, p. 2116):

$$Div_c = (n_c / N) \cdot GINI \cdot \left[\sum_{\substack{i=1 \\ j=1 \\ i \neq j}}^{\substack{i=n_c \\ j=n_c}} \frac{d_{ij}}{(n_c \cdot (n_c - 1))} \right], \tag{2}$$

where n_c is the number of categories, N the total number of categories, GINI is the Gini-coefficient, d_{ij} is the disparity the Euclidean distance in terms of 1-cosine (Ahlgren, Jarneving, & Rousseau, 2003) varying between 0 and 1, which is normalized by the factor (n_c (n_c-1)).

In this paper, a different path shall be taken, namely a path back to the information-theoretical roots of this concept, which is based on probabilistic foundations as already used by Shannon (1948). Although the initial work of Shannon and the entropy concept is mentioned in most publications and is also part of Stirling's concept, the probabilistic foundation and the classical entropy concept seems to have been lost or is currently not sufficiently received. For example, in a paper in Scientometrics, Shaw Jr. (1981) derives his concept of information theory from a thermodynamic representation of entropy, in which probabilities are only implicitly used. But why is probability theory for a consistent diversity concept so important? Four main reasons can be given:

1. *Unclear theoretical basis*: The theoretical basis for combining the different elements of diversity is not immediately apparent. For example, probabilities as p_j or p_i are linked to correlations or similarity measures (d_{ij}, Eq. 1, 2), which is not statistically derived from any probabilistic or statistical framework. Even Rao (1982, p. 7) himself says about the choice of d_{ij}: "The choice of d_{ij} is not a statistical problem and will depend on an individual's assessment of differences between qualitative categories reference to a given problem. However, one can use methods of multidimensional scaling in estimating d_{ij} by using *supplementary information* such as inequality relationships between d_{ij} and d_{rs} for different (i, j) and (r, s)."
2. *Accuracy*: Most diversity indicators are defined numerically, although the relative frequencies or probabilities are statistical quantities and thus have a statistical measurement error. Probabilities based on small sample sizes are far less accurate than probabilities estimated from large samples.

3. *Interpretation of indicators*: Given the large number of diversity indicators, the question remains as to how to interpret the individual indicators and what advantages one indicator has over another indicator.
4. *Statistical analysis*: Diversity indicators often have to be processed statistically, which in turn requires distributional assumptions. Therefore, the question is whether the data should not be formulated in a probabilistic form in order to apply the statistical analysis more or less directly to the raw data.

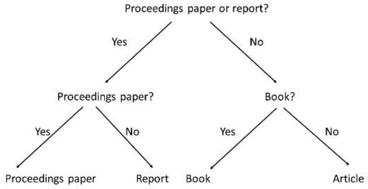
The central idea of the paper is to refer back to Shannon's concept of entropy, which is based on probability theory and forms the basis of today's diversity concepts, to arrive at a modified concept of diversity that firstly avoids previous inconsistencies in the diversity concept, secondly allows a simple form of interpretation in terms of uncertainty and reduction of uncertainty (entropy), and thirdly forms the basis both for new types of diversity indicators and for statistical modelling. In the following, both the modified diversity concept and an indicator will be derived in probabilistic terms, in order to then illustrate the approach with a practical application on research output data from Metrics Tide (https://responsiblemetrics.org/the-metric-tide/). The statistical model implementation must be, unfortunately, omitted here due to space constraints.

Probabilistic framework of information theory and modified diversity indicator div_e

Within information theory, information is defined purely syntactically without reference to semantics based on probability theory. In the following, a simple example will be taken as a starting point: The diversity of research output types of funded projects in the field of Natural Science (e.g., Mutz, Bornmann, & Daniel, 2012, 2014). Let us assume that project reports from 1000 funded projects are available as part of an expost evaluation (synthetic data). The frequencies of N=4 different output types are shown in Table 1, on the one hand before the data analysis without any information (prior or expectations) and on the other hand after the empirical data analysis (posterior).

Table 1. Relative frequencies for different output types (data example) and decision tree

Events	Output type	Code	prior	posterior
1	Books	00	0.25	0.10
2	Articles	10	0.25	0.80
3	Proceedings paper	01	0.25	0.10
4	Reports	11	0.25	0
Entropy			2	0.92



The following metaphor can be used to define information (Amann & Müller-Herold, 2011, p. 1f). If there is in advance no information about an event (here about the research output type), it is still possible to formulate a limited number of binary questions (Table 1, decision tree) to exactly determine which publication output is involved assuming the same probability of each event (Table 1, prior). The self-information is thus $\log_2(N)$, i.e. $\log_2(4) = 2$ bits (see Table 1, code). Another definition of self-information is obtained from probability theory, where the output type is a random variable X with j=1 to N possible events or occurrences (books, articles, ...). Thus, the self-information is $\log_2(1/p_j)$ or $-\log_2(p_j)$ with $p_j = 1/N$ if all events have the same probability (assumption $\log_2(p_j=0) = 0$). The average information over all events is then:

$$H(X) = \sum_{j=1}^{N} -p_j \log_2(p_j) \tag{3}$$

and is called Shannon entropy with the assumptions that $\sum_{j=1}^{N} p_j = 1.0$ and $\log_2(p_j=0) = 0$. It is assumed in a first step that the probability of occurrence of one event does not depend on the probability of occurrence of another event, an assumption which will be given up later. In research projects, the publication of proceedings papers should not depend on whether or not a journal article has also been published.

Shannon Entropy, as a measure of information, ultimately expresses the degree of uncertainty. For example, in a coin toss the uncertainty is highest when the probability of heads or tails is 0.5. A probability of 0.8 for heads, for example, reduces the uncertainty (tampered coin toss). People who bet on "heads" have a greater chance of winning than those who bet on "tails". In the above example, maximum entropy (H(X)=2) is reached when there is equal probability ($p_j = 0.25$) of the events. However, the actual observed frequencies as an estimate of the probabilities are not equal, the resulting entropy and thus uncertainty is reduced by 54% from 2 to 0.92 bits (see Table 1). This approach conforms to Bayes statistics, in which quasi in the context of an empirical learning process the initial expectation or uncertainty about model parameters (priors) is reduced (posterior) in the light of the data (Kruschke, 2011, p. 55). With the entropy concept, a first measure of diversity can be derived, that is "*balance*". For example, the more the observed probabilities resemble an equal probability, the higher is the so-called balance, the higher is the entropy or uncertainty.

Shannon also refers to "*variety*": "With equally likely events there is more choice, or uncertainty, when there are more possible events." (Shannon, 1948, p. 10). Furthermore, variety could also be traced back to an entropy measure. The maximum possible variety would be N, i.e. the maximum possible number of events (here N=4 output types). Finally, the maximum variety can be defined as the maximum entropy of the system as follows:

$$H_{Variety_{\max}} = \sum_{j=1}^{N} 1/N \log_2(1/N) = -\log_2(1/N) \tag{4}$$

In the case of four output types, the maximum variety would be $-\log_2(1/4)$. The observed variety corresponds to the number of events with nonzero probability $p_j > 0$:

$$H_{Variety_{obs}} = \sum_{j=1}^{N} 1/\sum_{j=1}^{N}(p_j > 0)\log_2(1/\sum_{j=1}^{N}(p_j > 0)) = -\log_2(1/\sum_{j=1}^{N}(p_j > 0)) \tag{5}$$

For the above example, the variety$_{obs}$ is $-\log_2(1/3)$. Similarly, Leydesdorff (2018, p. 2115) argues when defining a relative variety $\sum_{j=1}^{N}(p_j > 0)/N$. For an common diversity indicator one could add the two entropy masses balance and diversity, although this sum does not represent a joint-distribution H(A, B) in probabilistic terms, but can still be interpreted in terms of entropy with balance and variety as independent quantities:

$$H_{diversity} = H_{\text{var}iety} + H_{balance} \tag{6}$$

To get to the last component of diversity, "*disparity*", the assumption of independent events has to be abandoned. For example, research outputs in a research project may be published in different output types and this may also create stochastic dependencies. For example, the probability of a research article may depend on whether or not a proceedings paper has been published. In the simplest case, results of a research project can be published in two research outputs with a maximum of two different output types, where it is also possible that none of the output types is published or the same output type can be chosen twice. These two research outputs can be defined as two random variables X and Y.

Table 2 presents a cross-tabulation of cell frequencies and marginal frequencies of the two outputs. It quickly becomes clear that output 2 does not depend on output 1; the ratios of the relative frequencies across the columns remain the same across the rows. Thus p(X|Y) = P(X) or H(X, Y) = H(X) + H(Y), the marginal frequencies are sufficient ("overall balance").

Table 2. Case I: Independence

	Event	Output type	X (Output 2)				Total
			Books	Articles	Proceedings paper	Reports	
Y (Output 1)	1	Books	.01	.08	.01	.00	.10
	2	Articles	.08	.64	.08	.00	.80
	3	Proceedings paper	.01	.08	.01	.00	.10
	4	Reports	.00	.00	.00	.00	.00
	Total		.10	.80	.10	.00	1.00

For Table 2, the situation is different: P(X|Y) ≠ P(X), stochastic dependence is present. For example, articles in both outputs occur more frequently (p=.80) than expected in the case of independence of X and Y (p=.80*.80=.64).

Table 3. Case II: Dependence

	Event	Output type	X (Output 2)				Total
			Books	Articles	Proceedings paper	Reports	
Y (Output 1)	1	Books	.05	.00	.05	.00	.10
	2	Articles	.00	.80	.00	.00	.80
	3	Proceedings paper	.05	.00	.05	.00	.10
	4	Reports	.00	.00	.00	.00	.00
	Total		.10	.80	.10	.00	1.00

In that case, we speak of *mutual information* I(X, Y), i.e. the reduction of the uncertainty of one random variable by considering another random variable. In the case of two random variables, the mutual information is (Eshima, 2020, p. 8f):

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y \mid X) \qquad (7)$$

$$H(X \mid Y) = H(X) - I(X,Y)$$

The joint entropy of X and Y is then:

$$H(X,Y) = H(X) + H(Y) - I(X,Y) \qquad (8)$$

The mutual information I(X, Y) can be calculated form the data using the following equation:

$$I(X,Y) = \sum_{x\varepsilon X}\sum_{y\varepsilon Y} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right), \qquad (9)$$

where p(x, y) is the joint probability of X and Y. If the events are independent, then p(x, y) = p(x)p(y) and I(X, Y) =0. *The more the event Y depends on X, the higher the mutual information I(X, Y), the lower the total entropy H(X, Y).* Mutual information represents the stochastic dependence of X and Y and is not a correlation. In Figure 2, the relationships of the different concepts of entropy and mutual information are shown as a Venn diagram.

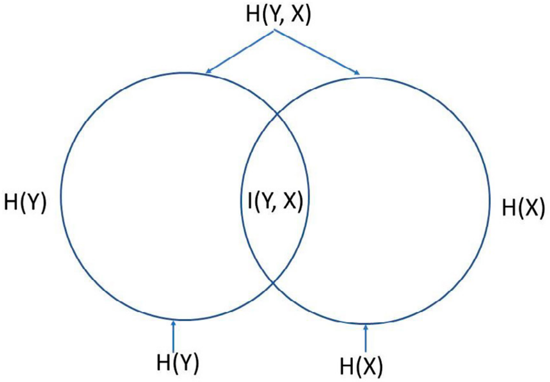


Figure 1. Entropy measures

Finally, mutual information, already mentioned by Shannon (1948, p. 12), defines the third component of diversity, "*disparity*", i.e. the degree of dependence of categories or events. The higher I(X, Y), the lower the disparity. If I(X, Y) is zero, then the disparity is maximal.

In the above example (Table 3), the mutual information is I(X, Y)=.72, the unconditional entropy of X is H(X)=.92, and the conditional entropy is H(X|Y)=.20. This reduces the uncertainty, i.e. overall diversity, by (.92-.20)/.92)=78.3% when taking into account the dependence in the data in the sense of disparity.

In the case of 3 research outputs with a maximum of 3 different output types, for example, the above formula, corresponding to the addition theorem of probabilities, expands as follows:

$$H(Y,X,Z)=H(Y)+H(X)+H(Z)-I(Y,X)-I(Y,Z)-I(X,Z)+I(Y,X,Z)) \qquad (10)$$

Where the last term of Eq. 10 is positive and represents a correction term in a quantity algebraic representation (i.e. subtraction of an intersection).

Unlike Stirling and Leydesdorff (Leydesdorff, 2018; Leydesdorff, Wagner, & Bornmann, 2019; Stirling, 2007), the different components of the *overall diversity indicator div_e* are now additive:

$$div_e = H_{\mathrm{var}iety} + H_{balance} - I_{disparity}\,, \qquad (11)$$

where div_e=4.15 for the above example.

In summary: On the probabilistic basis of Shannon's entropy concept, a modified diversity concept was derived that is consistent, additive in nature and whose components can be empirically derived. The entropy concept gives the diversity concept a clear interpretation: the higher the diversity, the higher the information, the higher the uncertainty, the less structure there is in the data.

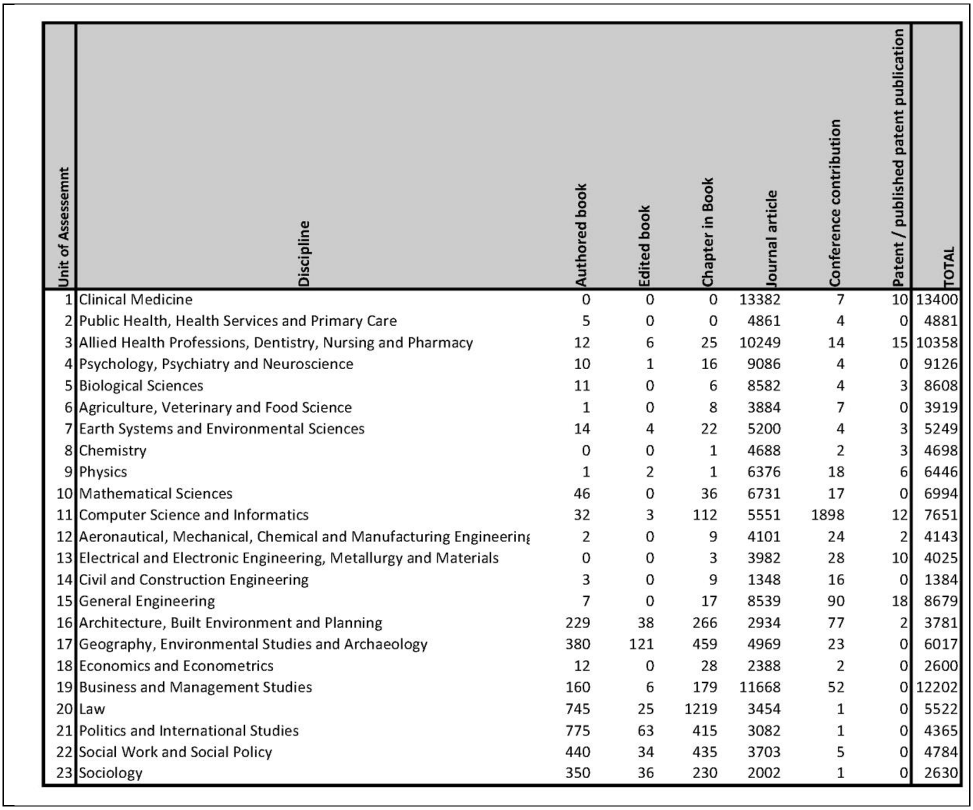
Data and methods

To answer the question of how diverse research outputs are across research areas, we drew on data from the Metric Tide Report. Metric Tide analyzed the capabilities and limitations of research metrics and indicators. "It has explored the use of metrics across different disciplines, and assessed their potential contribution to the development of research excellence and impact.

It has analyzed their role in processes of research assessment, including the next cycle of the Research Excellence Framework (REF).

For the analysis, data on submissions for research projects funded since 2006 (Research Excellence Framework 2014) was used. Figure 2 shows an extract of the original cross table indicating how frequently 20 output types occur in 36 units of assessments (scientific disciplines). It would have been desirable for the analysis to have data on the output types for each individual research project. Such raw data were not available. Furthermore, it is not clear whether the frequencies represent the number of research projects in a cell or the sum of research outputs across all research projects in that cell.

To arrive at relative frequencies or probabilities, the absolute frequencies were divided by the total number (N = 190.962). Random variable X is the "output type", and random variable Y is the "research area" (unit of assessment, UOA).



Unit of Assessemnt	Discipline	Authored book	Edited book	Chapter in Book	Journal article	Conference contribution	Patent / published patent publication	TOTAL
1	Clinical Medicine	0	0	0	13382	7	10	13400
2	Public Health, Health Services and Primary Care	5	0	0	4861	4	0	4881
3	Allied Health Professions, Dentistry, Nursing and Pharmacy	12	6	25	10249	14	15	10358
4	Psychology, Psychiatry and Neuroscience	10	1	16	9086	4	0	9126
5	Biological Sciences	11	0	6	8582	4	3	8608
6	Agriculture, Veterinary and Food Science	1	0	8	3884	7	0	3919
7	Earth Systems and Environmental Sciences	14	4	22	5200	4	3	5249
8	Chemistry	0	0	1	4688	2	3	4698
9	Physics	1	2	1	6376	18	6	6446
10	Mathematical Sciences	46	0	36	6731	17	0	6994
11	Computer Science and Informatics	32	3	112	5551	1898	12	7651
12	Aeronautical, Mechanical, Chemical and Manufacturing Engineerin	2	0	9	4101	24	2	4143
13	Electrical and Electronic Engineering, Metallurgy and Materials	0	0	3	3982	28	10	4025
14	Civil and Construction Engineering	3	0	9	1348	16	0	1384
15	General Engineering	7	0	17	8539	90	18	8679
16	Architecture, Built Environment and Planning	229	38	266	2934	77	2	3781
17	Geography, Environmental Studies and Archaeology	380	121	459	4969	23	0	6017
18	Economics and Econometrics	12	0	28	2388	2	0	2600
19	Business and Management Studies	160	6	179	11668	52	0	12202
20	Law	745	25	1219	3454	1	0	5522
21	Politics and International Studies	775	63	415	3082	1	0	4365
22	Social Work and Social Policy	440	34	435	3703	5	0	4784
23	Sociology	350	36	230	2002	1	0	2630

Figure 2. Extract of the table of frequencies for 23 research areas (UOA) × 6 output types (Wilsdon et al., 2015, p. 154)

Results

Different measures can be calculated for the data (Table 4). Thus, a statistically significant χ^2-test value, a Cramer`s V of .24 and a mutual information of 0.40 show that the output type (X) depends on the research area (Y). There are, as expected, significant differences of the research area in the output types of the research output beyond chance. Looking at the ratio of the unconditional entropy of "output type", H(X), and "research area", H(Y), to the total entropy, H(X, Y), it is clear that the differences in frequencies are very much determined by the differences in "research areas" than the differences between "output types". While considering the "research area" reduces about (1.20-0.80)/1.20=33% of the uncertainty in the "output types", conversely only about (4.96-4.56)/4.96=8% of uncertainty in the "research areas" is reduced.. When the balances H(X) or H(Y) are compared with the respective maximum balances, it becomes clear that the balance in the "output types" of H(X) = 1.197 compared to H_{max}(X) of 4.322 is much lower than in the "research areas" (H(Y)=4.961, H_{max}(Y)=5.170). Research output types are essential less balanced than the research areas.

Overall, however, the diversity at 10.88 is very high, which is about 78.8% of the maximum possible diversity of 13.81. The maximal variety in the whole table is fully reached and the overall balance of 6.16 is 64.9% of the maximum possible balance of 9.49.

Table 4. χ^2 and entropy measures

Measure	Label	Value	Maximum
χ^2	χ^2-test value	204,768.3*	
Cramer`s V	Cramer`s V (correlation)	0.24	1.00
I(X, Y)	Mutual information ("disparity")	0.40	0.00
H(X)	Uncond. entropy "output type" ("balance")	1.20	4.32
H(X|Y)	Condit. entropy "output type"	0.80	
H(Y)	Uncond. entropy "research area" ("balance")	4.96	5.17
H(Y|X)	Condit. entropy "research area"	4.56	
H(X, Y)	Total entropy	5.76	9.49
Overall indices			
$H_{variety}$	Variety	4.32	4.32
$H_{balance}$	Balance	6.16	9.49
$H_{disparity}$	Disparity	0.40	0
div_e	Diversity	10.88	13.81

* p<.05 (df=665)

Table 5. Diversity measures for a set of 11 research areas (UOA)

UOA	Name	N	Variety	Balance	Disparity	Diversity
1	Clinical Medicine	4	2.00	0.27	0.02	2.25
2	Public Health, Health Services and Primary Care	4	2.00	0.14	0.01	2.13
3	Allied Health Professions, Dentistry, Nursing and Pharmacy	8	3.00	0.23	0.01	3.22
4	Psychology, Psychiatry and Neuroscience	9	3.17	0.21	0.01	3.37
5	Biological Sciences	7	2.81	0.20	0.01	3.00
6	Agriculture, Veterinary and Food Science	6	2.58	0.12	0.01	2.70
7	Earth Systems and Environmental Sciences	7	2.81	0.15	0.01	2.95
8	Chemistry	5	2.32	0.13	0.01	2.45
9	Physics	8	3.00	0.17	0.01	3.16
10	Mathematical Sciences	9	3.17	0.19	0.01	3.35
11	Computer Science and Informatics	16	4.00	0.23	0.03	4.19

Note. N=number of different output types $\Sigma(p_j>0)$

Finally, diversity measures can be derived for the "research areas", which provide information about the diversity for individual research areas. Diversity indicators for the first 11 research areas are shown in Table 5. “Computer Science and Informatics" has the highest diversity with 4.19 and "Public Health, Health Services and Primary Care" the lowest with 2.13.
"Computer Science and Informatics" has the highest variety of 4.0. "Clinical Medicine" has the highest balance (.27) in this set of selected research areas.
Among the different output types, "Journal articles" clearly shows the highest individual entropy, H(Y), i.e. the highest balance among all other output types across different research area (Table 6).

Table 6. Output types with highest individual balance values

Output type	Balance
Journal article	4.15
Chapter in book	0.59
Authored book	0.45
Conference contribution	0.12
Edited book	0.11
Exhibition	0.05

Discussion

Diversity is a ubiquitous term used in many disciplines (e.g., ecology, sociology, bibliometrics). A certain break in the discussion on diversity and diversity indicators was brought about by the indicator developed by Rao (1982) and Stirling (2007), which in view of its comparatively simple definition has a very wide circulation. It is to the credit of Rousseau (2019) to point out inconsistencies in this indicator, to which Leydesdorff (2018) has proposed a workable solution. Furthermore, statistical concepts such as probability and correlation / similarity are combined with each other in a way for which there is no statistical basis whatsoever, which was even noted by Rao (1982, p. 7).

Table 7. Data example for the calculation of the diversity indicator *div_e*

	Research Output			
Project-ID	Journal articles	Books	Proceeding papers	Bookchapter
1	1	0	1	0
2	1	1	0	1
3	0	1	0	1
…				

Note. 1 = occurred in the project, 0 = not occurred in the project.

Due to the problems of the Rao and Stirling indicator, the aim of this paper was to go back to Shannon's probabilistic concept of entropy (Shannon, 1948), which implicitly dealt with all three facets of diversity, in order to develop a modified concept of diversity from it, which is additive in its nature and allows for both the calculation of a diversity indicator div_e as well as estimation diversity by a statistical model. The typical data structure is like that in Table 7.
In the sense of Occam's razor principle, the question arises as to why this more complex approach should be given preference in practice over the simple concept of Stirling and Rao. First of all, there is no intention to replace the existing diversity indicators, but to identify opportunities for improvement. The following four reasons could be put forward:

1. *Inconsistencies*: Concepts with obvious inconsistencies are not very convincing and reinforce the negative image of bibliometrics.

2. *Interpretation*: Diversity can be interpreted in terms of entropy as a measure of information. Diversity is at its maximum when events can no longer be predicted, as in a coin toss. The lower the diversity, the more predictable events are, the more structure is in the data.
3. *Open the Pandorra`s box*: The discussions on bibliometric concepts such as field normalization, definition of fields, fractional counting and also diversity seem to be more or less closed with some more or less workable solutions. These considerations might reopen the discussion.
4. *Stochastic nature*: If the occurrence of publications, research output, citations, etc. is assumed to base on a stochastic random process, this must be taken into account within the development of an indicator.
5. *Statistical estimation*: In principle, diversity and its various components can be in principle statistically estimated, which is the subject of future publications (e.g., Bornmann, Mutz, & Daniel, 2013; Bornmann, Stefaner, de Moya Anegón, & Mutz, 2016).

In the end, the present contribution could, actually, only represent an idea of an indicator. There is still a lack of further analysis showing the behavior of the indicator in comparison to other indicators. Furthermore, the calculation of the disparity indicator requires a workable solution on how to deal with the number of components that increases with the number of units (e.g. research output types). For example, with 3 units, there are 4 components of the disparity indicators. Last but not last, efficient statistical models must be developed to estimate diversity as a statistical parameter.

Acknowledgement

Figure 2 is an extract of a table from Wilsdon et al. (2015), which is in part copyright HEFCE under Open Government Licence v2.0.

References

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology, 54*(6), 550-560. doi: https://doi.org/10.1002/asi.10242

Amann, A., & Müller-Herold, U. (2011). *Offene Quantensysteme [Open Quants systems]*. Berlin: Springer.

Bornmann, L., Mutz, R., & Daniel, H. D. (2013). Multilevel-statistical reformulation of citation-based university rankings: The Leiden ranking 2011/2012. *Journal of the American Society for Information Science and Technology, 64*(8), 1649-1658. doi:10.1002/asi.22857

Bornmann, L., Stefaner, M., de Moya Anegón, F., & Mutz, R. (2016). Excellence networks in science: A Web-based application based on Bayesian multilevel logistic regression (BMLR) for the identification of institutions collaborating successfully. *Journal of Informetrics, 10*(1), 312-327. doi:10.1016/j.joi.2016.01.005

Eshima, N. (2020). *Statistical data analysis and entropy*. Singapore: Springer Nature.

Goyanes, M., Demeter, M., Grané, A., Albarrán-Lozano, I., & Gil de Zúñiga, H. (2020). A mathematical approach to assess research diversity: operationalization and applicability in communication sciences, political science, and beyond. *Scientometrics, 125*(3), 2299-2322. doi:10.1007/s11192-020-03680-6

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology, 88*(10), 2427-2439. doi:10.1890/06-1736.1

Kruschke, J. K. (2011). *Doing Bayesian data analysis - A tutorial with R and BUGS*. Burlington: Elsevier.

Leydesdorff, L. (2018). Diversity and interdisciplinarity: how can one distinguish and recombine disparity, variety, and balance? *Scientometrics, 116*(3), 2113-2121. doi:10.1007/s11192-018-2810-y

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Diversity measurement: Steps towards the measurement of interdisciplinarity? *Journal of Informetrics, 13*(3), 904-905. doi:https://doi.org/10.1016/j.joi.2019.03.016

Mutz, R., Bornmann, L., & Daniel, H.-D. (2012). Types of research output profiles: A multilevel latent class analysis of the Austrian Science Fund's final project report data. *Research Evaluation, 22*(2), 118-133. doi:10.1093/reseval/rvs038

Mutz, R., Bornmann, L., & Daniel, H.-D. (2014). Cross-disciplinary research: What configurations of fields of science are found in grant proposals today? *Research Evaluation, 24*(1), 30-36. doi:10.1093/reseval/rvu023

Rao, R. R. (1982). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 44*(1), 1-22.

Rousseau, R. (2018). The repeat rate: from Hirschman to Stirling. *Scientometrics, 116*(1), 645-653. doi:https://doi.org/10.1007/s11192-018-2724-8

Rousseau, R. (2019). On the Leydesdorff-Wagner-Bornmann proposal for diversity measurement. *Journal of Informetrics, 13*(3), 906-907. doi:https://doi.org/10.1016/j.joi.2019.03.015

Rousseau, R., Van Hecke, P., Nijssen, D., & Bogaert, J. (1999). The relationship between diversity profiles, evenness and species richness based on partial ordering. *Environmental and Ecological Statistics, 6*(2), 211-223. doi:10.1023/A:1009626406418

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(4), 379–423. doi:doi:10.1002/j.1538-7305.1948.tb01338.x

Shaw Jr., W. M. (1981). Information theory and scientific communication. *Scientometrics,, 3*(3), 235-249. doi:https://doi.org/10.1007/BF02101668

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707-719. doi:10.1098/rsif.2007.0213

Wang, J., Thijs, B., & Glänzel, W. (2015). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PLoS ONE, 10*(5). doi:10.1371/journal.pone.0127298

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., . . . Johnson, B. (2015). *The Metric Tide: Report on the independent review of the role of metrics in research assesment and management*. Retrieved from https://responsiblemetrics.org/wp-content/uploads/2019/02/2015_metrictide.pdf

Datasets on DataCite - an Initial Bibliometric Investigation

Anton Ninkov[1,2], Kathleen Gregory[1,2,4], Isabella Peters[3,5] and Stefanie Haustein[1,2,6]

[1] School of Information Studies, University of Ottawa, Ottawa (Canada)
[2] Scholarly Communications Lab, Ottawa/Vancouver (Canada)
[3] ZBW Leibniz Information Center for Economics, Kiel (Germany)
[4] Data Archiving & Networked Services, Royal Netherlands Academy of Arts & Sciences (Netherlands)
[5] Kiel University, Kiel (Germany)
[6] Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, Montreal (Canada)

Abstract

Interest in measuring data citation and developing metrics for data is increasing. Despite this interest, basic bibliometric research investigating data sharing, data reuse and data citation practices remains relatively nascent. In this research in progress article, we use the DataCite GraphQL API to gather data for an initial investigation into dataset sharing and reuse as well as consider the current challenges. With over 8 million datasets in DataCite, we look at how datasets are dispersed by publication year, discipline, number of citations, license, institutional affiliation, and language. We find some patterns emerging, such as a recent increase in dataset publishing. However, there are still many limitations to doing this research that are discussed. As well, the future use of DataCite as a resource for doing this research and additional methods of analysis are considered.

Introduction

As data are increasingly becoming recognized scholarly outputs, funders, research managers and publishers are interested in developing data metrics to reflect the usefulness and impact of sharing data (Cousijn et al., 2019). Despite the interest in metrics for data, basic bibliometric research investigating data sharing, data reuse and data citation practices remains an underserved area.

The obstacles for conducting bibliometric research focusing on data are complex, involving decisions made in policies, practices and at the technical level (Borgman, 2016). These factors are compounded by a lack of bibliometric evidence about data sharing and reuse, particularly a lack of standardization in data citation practices. This creates a so-called "vicious circle," where bibliometricians tend not to take data as an object of study, while at the same time such research is required for developing meaningful data metrics and best practices in the field (Morissette, Peters, & Haustein, 2020).

This paper takes the first steps in addressing this vicious circle, presenting a preliminary bibliometric investigation into data sharing and citation practices, using metadata from DataCite (https://datacite.org) as a source. This research in progress article, which is part of a collaborative project involving DataCite and bibliometricians, provides an overview of the current state of the data available from DataCite. While perhaps not as large as other corpuses, DataCite is a relevant resource for bibliometric research as it is: a) not focused on a single discipline, and b) it assigns persistent identifiers (i.e. DOIs) to research data, allowing for a robust tracking of citations. Given documented disciplinary differences in data sharing, reuse and citation practices (Borgman, 2015; Tenopir et al., 2015), and the importance of accounting for disciplinary differences in data metric development (Lowenberg et al., 2019), we pay special attention to the presence (or absence) of information in DataCite about disciplinary domains in our analysis. We conclude our analysis by identifying gaps in the available data from DataCite and highlighting future areas for data-centric bibliometric research.

Background

Data sharing and citation

Data citation, and calls for its standardization, are not new matters of concern (Parsons et al., 2019). Milestones in the development of standards include the Bermuda Principles in 1996, the formation of CrossRef in 1999, and the founding of DataCite in 2009 (Lowenberg et al., 2019). Silvello (2018) extensively analyzes the extant literature on the development of such standards, as well as motivations for data citation current technical systems. The MDC initiative and DataCite have particularly contributed to efforts on the standardization of data citations in recent years (i.e. Fenner et al., 2019), as has the Scholix Framework for Interoperability in Data-Literature Information Exchange (Burton et al., 2017) and recommendations developed within the Research Data Alliance (Rauber et al., 2015). However, these projects focus on data citation infrastructure, not bibliometric research on data citation practices.

At a more granular level, other work analyses the state of data sharing, reuse and citation for individual datasets. Such work highlights the impermanent and untrackable nature of some citations, such as the widespread use of URLs to reference data (Yoon et al., 2019), or the practice of including data references in the body of articles or in acknowledgement sections, rather than in reference lists (Park et al., 2018). A general laissez-faire approach to data citation has been noted in a number of other studies (Fecher et al., 2015), and persists even in cases where recommended citation formats are provided (Belter, 2014).

Disciplinary differences in data sharing and citation remain a recognized yet unsolved problem. Disciplinary norms play an important factor in the willingness to share datasets (Tenopir et al., 2015). Similarly, early work analyzing the Thomson Reuters (now Clarivate) Data Citation Index (DCI) finds that the hard sciences, specifically biomedical fields, account for the majority (80%) of entries (Torres-Salinas et al., 2014). Lowenberg et al. (2019) note that disciplinary differences in the conception of data themselves demand creating discipline-specific usage statistics (i.e., downloads, views). Peters et al. (2016) also caution that statistics derived from data citations must be interpreted in the context of (disciplinary) data sharing practices and norms, finding that 85% of data remain uncited.

DataCite

The case study described in this research in progress article uses DataCite as the source for our bibliometric investigation. DataCite is an international, non-profit organization that has been assigning persistent identifiers, DOIs, for research data and other artefacts since 2009. DataCite is also actively involved in community outreach and provides data management services and support. Institutions become DataCite members to obtain DOIs for their resources. To receive a DOI, members provide DataCite with metadata describing the given data or artefact. This metadata is provided according to a specialized schema consisting of optional, recommended and mandatory fields (see Table 1).

Table 1. Mandatory Fields by DataCite

Field	*Description*
Identifier	The Identifier is a unique string that identifies a resource.
Creator	The main researchers involved in producing the data, or the authors of the publication, in priority order.
Title	A name or title by which a resource is known.
Publisher	The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource.
Publication Year	The year when the data was or will be made publicly available.
Resource Type	A description of the resource.

DataCite also collects citation data that can be accessed through the GraphQL API. There are two ways in which DataCite collects this data: DataCite members provide the information or DataCite learns about the citation from other academic resources, i.e. CrossRef (Garza, 2020). Recent studies examine DataCite's coverage and potential role in the development of scholarly metrics. Robinson-Garcia et al. (2017), for example, highlight the lack of a standardized vocabulary amongst metadata field entries as a hindrance to its utility as a metrics source. Another recent report examining the role of DataCite in open science practices supports these findings (Dudeck et al., 2019). Using a sample of datasets from an ocean science repository, the authors further conclude that data reuse within this sample is limited to a small number of organizations or to reuse by the original data creators. The study we present here fits into this literature, taking a high-level approach to provide a current analysis of the state of data within DataCite. We consider features that have not been examined before (e.g., citation data) presenting an initial step to more detailed future analyses.

Methodology

To gather data on datasets, we have used DataCite's GraphQL API to collect the metadata. Data was collected between April 15, 2021. Because DataCite is constantly having new submissions and uploads, it was important to collect data in a short time period. With the GraphQL API queries saved, future research collecting data to investigate any rapid changes to DataCite can be done quickly.

One important limitation to mention is that at the time of data collection, the DataCite GraphQL API only returned a maximum of 10 items per query. This is a design feature of the GraphQL API to help optimize performance. Because of this 10-item limitation, accessing the large amounts of data that is offered by DataCite is a challenge, and required work arounds both in terms of methodology and research questions (e.g., running multiple specified queries). There are ways, however, to customize queries in accordance with the DataCite team which can be explored in future work.

Initial Findings

At the time of data collection, there are 8,643,593 total datasets indexed in DataCite (7,440,415 records in 2017; Robinson-Garcia et al., 2017). The most frequent language of datasets is English, with more than 50% of all datasets in DataCite classified as such. Of the ten most common languages of datasets, all but one (Thai) are European languages. Of all the datasets found in DataCite, the majority have been published in the last 10 years (2011-2020). This accounts for 86% of the total number of datasets. It should be noted that these dates are for when the data was published, and not when the data was collected or used. In Table 2, the number of published datasets in DataCite by decade are listed and the general trend of a recent increase is noticeable. One observation is that there is a steep increase 1931-1940 from the previous and following decade. This is a result of a specific repository (University College Dublin) uploading a batch of data during this time period. This repository had a great deal of data connected to this time period as a result of a specific project covering 1937-1938.

Table 2. DataCite Datasets Published by Decade

Decade	*Number of Dataset*	*Decade*	*Number of Datasets*
2011-2020	7,129,806	1961-1970	8,766
2001-2010	815,399	1951-1960	544
1991-2000	87,782	1941-1950	212
1981-1990	30,249	1931-1940	48,073
1971-1980	10,694	1921-1930	53

There are currently 535,449 datasets in DataCite that have a discipline specified, ca. 6% of all datasets. This is not enough data to do a thorough investigation of discipline behaviors when it comes to data sharing and citation practices. However, based on personal communication with technical experts at DataCite, there will be over four million datasets that will gain discipline classification in 2021. This will allow for a more robust study of this area. Of the datasets that have a discipline identified, the ten disciplines with the most published datasets are listed in Table 3. For each discipline, the number of datasets with a citation are also listed. It should be noted that there are very few datasets in DataCite that currently have citations. In total, 97,734 of the datasets have at least one citation, ca. 1% of all datasets.

Table 3. Ten Most Common Disciplines Listed for DataCite Datasets

Discipline	*Number of Datasets*	*Number of Datasets with ≥1 Citation*
Biological sciences	289,137	286
Earth + related environmental sciences	89,931	414
Health sciences	77,601	29
Chemical sciences	63,285	9
Computer + information sciences	61,483	52
Clinical medicine	58,702	33
Sociology	39,144	166
Mathematics	32,901	12
Physical sciences	17,660	24
Psychology	15,450	25

The size of the dataset is listed for 2,485,517 datasets, or 28% of the total number of datasets. The reporting of the data for this variable is not standardized which makes its analysis challenging. Authors of datasets input information on size free of formatting. For example, some datasets express size by how many bytes a dataset takes up (e.g. 2GB) while others record this in other ways (e.g., number of rows, number of items). With the lack of standardization for size, there is a limited amount of data that can be provided by DataCite and, therefore, the depth of analysis we can conduct.

A mandatory field when inputting a dataset into DataCite is publisher. This variable, like size, does not have standardized data input, which makes evaluation of the metadata challenging and not something that is available via the GraphQL API. However, there are some general observations that can be made based on internal DataCite data. Global Biodiversity Information Facility is currently the most frequently identified publisher (941,335 datasets) and The Cambridge Structural Database is the second most frequent (889,586). There is a discrepancy between our findings and those reported in Robinson-Garcia et al. (2017) that requires further investigation in future work. Again, with the lack of standardization, there is a limitation to the data provided by DataCite and the analysis that we can conduct.

The license that is assigned to a dataset is important because it has a direct effect on the options to reuse a dataset. In Table 4, we have listed the top 10 most common licenses for datasets in DataCite. These licenses have been listed in order of the year in which the license was most used, and the number of datasets for that year and total datasets are listed. Because there has been such a recent uptick in use of DataCite, it is not that surprising that so many of the licenses have been published 2020. It is interesting to note how CC-BY-3.0 was so frequently used in 2006. This license was released in 2007 which would explain why it would have been so heavily adopted by these datasets.

Table 4. DataCite Datasets Published by License

License	*Year of Most Published*	*Number of Datasets for Most Published Year*	*Number of Total Datasets*
CC-BY-4.0	2020	171,807	625,685
CC-BY-NC-4.0	2020	102,918	214,311
CC-BY-NC-SA-4.0	2020	9,487	15,630
CC-BY-SA-4.0	2020	2,776	5,586
CC-BY-NC-3.0	2020	1,314	4,123
CC-BY-NC-ND-4.0	2020	634	2,378
CC-BY-1.0	2020	188	399
MIT	2018	1,036	4,281
CC0-1.0	2016	32,098	140,342
CC-BY-3.0	2006	55,848	283,489

Discussion

The goal of this research in progress article is to give initial insights into dataset reuse, availability, and sharing behaviors. To examine variables including datasets' age, size, license, publisher, language, citations, and discipline, we have used the DataCite GraphQL API. With that we contribute to the bibliometric meta research on datasets, hoping to encourage other bibliometricians to explore this type of scholarly output in more detail.

In bibliometric studies, discipline is an important scholarly variable which we have found to also need more attention in dataset studies. However, with only 6% of datasets in DataCite currently having a discipline classification, there is not enough data to do a robust analysis of datasets by discipline. More (approximately 50% of total) needed discipline classification will be added to DataCite in 2021. Other possibilities for expanding the amount of discipline data that is available will have to be considered, which could include integrating metadata from other sources to infer disciplines. As well, getting publishers and repositories to deliver data on disciplines in the future by asking them to select from controlled vocabulary (such as those provided by Organisation for Economic Co-operation and Development) could be necessary. We believe that this aspect is a rich field for future bibliometric research.

There are other variables, in addition to the ones presented in this research in progress article, that would be useful for this type of analysis. For example, additional variables on dataset reuse (e.g., downloads), the number of authors/contributors or the countries of origin could reveal insights into the nature of dataset sharing. There are challenges to doing this right now, mostly surrounding not having enough data to study it. With the increasing adoption of DataCite as well as DataCite adding more variables to their API we project that this will become an even more important area of research imminently.

Acknowledgments

The Alfred P. Sloan Foundation (Sloan Grant G-2020-12670) supported this work. We would like to thank Daniella Lowenberg (DL), Kristian Garza (KG) and Martin Fenner (MF). Stefanie Haustein, Isabella Peters, MF, DL, and KG collaborate in the 'Make Data Count initiative'.

References

Belter, C. W. (2014). Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. PLoS ONE, 9(3), e92590. https://doi.org/10.1371/journal.pone.0092590

Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world.

Borgman, C. L. (2016). Data citation as a bibliometric oxymoron. In C. R. Sugimoto (Ed.), *Theories of informetrics and scholarly communication* (pp. 93–116). De Gruyter. https://doi.org/10.1515/9783110308464-008

Burton, A., Aryani, A., Koers, H., Manghi, P., La Bruzzo, S., Stocker, M., Diepenbroek, M., Schindler, U., & Fenner, M. (2017). The Scholix Framework for Interoperability in Data-Literature Information Exchange. D-Lib Magazine, 23(1/2). https://doi.org/10.1045/january2017-burton

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & Simons, N. (2019). Bringing Citations and Usage Metrics Together to Make Data Count. *Data Science Journal*, *18*(1), 9. DOI: http://doi.org/10.5334/dsj-2019-009

Fecher, B., Friesike, S., & Hebing, M. (2015). What Drives Academic Data Sharing? PLOS ONE, 10(2), e0118053. https://doi.org/10.1371/journal.pone.0118053

Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., & Clark, T. (2019). A data citation roadmap for scholarly data repositories. Scientific Data, 6(1), 28. https://doi.org/10.1038/s41597-019-0031-8

Garza, K. (2020). Datacite Citation Display: Unlocking Data Citations. DataCite. https://doi.org/10.5438/1843-K679

Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., & Jones, M. B. (2019). Open Data Metrics: Lighting the Fire. Zenodo. https://doi.org/10.5281/zenodo.3525349

Morissette, Erica, Peters, Isabella, & Haustein, Stefanie. (2020). Research data and the academic reward system. Zenodo. http://doi.org/10.5281/zenodo.4034585

Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. Journal of the Association for Information Science and Technology, 69(11), 1346-1354.

Parsons, M. A., Duerr, R. E., & Jones, M. B. (2019). The History and Future of Data Citation in Practice. *Data Science Journal*, *18*, 52. https://doi.org/10.5334/dsj-2019-052

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: An extended analysis of citations and altmetrics. Scientometrics, 107(2), 723–744. https://doi.org/10.1007/s11192-016-1887-4

Rauber, A., Asmi, A., van Uytvanck, D., & Proell, S. (2015). Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). http://dx.doi.org/10.15497/RDA00016

Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. Journal of Informetrics, 11(3), 841–854. https://doi.org/10.1016/j.joi.2017.07.003

Silvello, G. (2018). Theory and practice of data citation. Journal of the Association for Information Science and Technology, 69(1), 6-20.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE, 10(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Torres-Salinas, D., Martín-Martín, A., & Fuente-Gutiérrez, E. (2014). Analysis of the coverage of the Data Citation Index – Thomson Reuters: Disciplines, document types and repositories. Revista Española de Documentación Científica, 37(1), e036. https://doi.org/10.3989/redc.2014.1.1114

Yoon, J., Chung, E., Lee, J. Y., & Kim, J. (2019). How research data is cited in scholarly literature: A case study of HINTS: How HINTS data is cited in scholarly literature. Learned Publishing, 32(3), 199–206. https://doi.org/10.1002/leap.1213

A Generalized Power Law Model of Citations

Ruheyan Nuermaimaiti[1], Leonid V. Bogachev[2] and Jochen Voss[3]

[1] *mmrn@leeds.ac.uk* [2] *L.V.Bogachevz@leeds.ac.uk* [3] *J.Voss@leeds.ac.uk*
School of Mathematics, University of Leeds, Leeds LS2 9JT (United Kingdom)

Abstract

The classical power law model is widely used in informetrics to describe citations of scientific papers, although it is not addressing variability across individual authors. We report our preliminary results for a novel model based on a certain parametric form of the expected individual citation profile which generalizes the power law frequency formula. The new model interpolates between large citation numbers, where the power law tail is reproduced, and low citation numbers, which are usually truncated when fitting the power law model to the data. In addition, we derive a deterministic limit shape of the citation profile, which can be used to make predictions about various citation function such as the h-index.

Introduction

The classical power-law model was introduced by Lotka (1926) as an empirical match with observed frequencies of citations in scientific publications. In a later development, Price (1965) discovered an important connection with networks, whereby citations were interpreted as nodes' degrees. Examples of fitting the power law to the citation data can be found in Coile (1977), Redner (1998), and Clauset, Shalizi & Newman (2009). In particular, it was found that the power law frequencies do not necessarily fit well in the entire citation spectrum, so that a suitable truncation of lower citation values may be needed.

Importantly, no assumptions are made in the power law model about the frequency distribution of citations for an individual author randomly chosen from the population of authors. This makes it difficult to project the model fitted to a pooled corpus of publications onto individual authors, for example, for the purposes of evaluating their productivity.

The power-law model can be fitted to real-life data using standard statistical methods such as the maximum likelihood or ordinary least squares estimation. As has been documented across many use cases (Clauset, Shalizi & Newman, 2009), the power law usually fits quite well but only in the tail region of the frequency range, which motivates the use of truncated power-law models by excluding the lower values. This may decrease the utility of the model in estimation of various functions of citations, such as the popular h-index, introduced by Hirsch (2005) and defined as the maximum number h of an author's papers, each one cited at least h times.

In an attempt to overcome this shortcoming, we propose a novel model by modifying the power law setting. The new model interpolates between slow (almost flat) decay of the citation frequencies at the bottom of the citation spectrum and then reproducing the power-law behavior at the tail of the frequency distribution. As we will demonstrate below using a small real data set, the model provides a very good fit across the entire citation spectrum. In addition, and in contrast to the scale-free power law, our model possesses a deterministic limit shape of the citation profile, which can be used, for example, to make meaningful estimation of the h-index. In particular, the estimation of the h-index based on the modified model appears to be significantly more accurate as compared to that in the standard power law model.

Power law frequencies

In its classical setting, the power law model states that the relative frequency f_j of exactly j citations accumulated by a randomly sampled paper is proportional to the a-th power of j, with some exponent $a > 1$ (typically lying in the range $2 < a < 3$), that is (to include the case $j = 0$ and to normalize the sum of frequencies to unity),

$$f_j = \frac{c}{(j+1)^a} \qquad (j = 0,1,2,\dots), \tag{1}$$

where the normalizing constant c is given by the reciprocal Riemann zeta function,

$$c^{-1} = \sum_{j\geq 0} \frac{1}{(j+1)^a} = \zeta(a). \tag{2}$$

Here, "randomly sampled" means that a paper is sampled from a *pooled corpus* of papers written by a (large) population of authors. For simplicity, interaction effects due to joint authorship are not taken into account; such effects are complicated but have minor impact.

In terms of analysis of citation data, if there are M_j papers with j citations, out of the total number of papers M, then the power law model predicts that the relative frequencies M_j/M are approximately given by formula (1),

$$f_j \approx \frac{M_j}{M} \qquad (j = 0,1,2,\dots). \tag{3}$$

Note that the total number of papers and the total number of citations in this corpus are given, respectively, by

$$M = \sum_{j\geq 0} M_j\,, \qquad N = \sum_{j\geq 0} j\, M_j. \tag{4}$$

Of course, in any real-life data set the numbers M_j will reduce to zero for j big enough (so that the series in (4) are in fact finite sums), but this is reconciled with the prediction (1) simply by the fact that the theoretical frequencies f_j tend to zero as $j \to \infty$.

As mentioned in the Introduction, the power law model does not address the frequency distribution of citations for an *individual author* randomly chosen from the observed population of authors (say, of size K) and featured by a collection of *citation counts* ν_j, that is, the numbers of papers by this author that have j citations ($j = 0,1,2, \dots$). Moreover, the number of observed authors, K, is often omitted in popular citation data sets (cf. Redner, 1998).

Having in mind statistically homogeneous populations of authors who produce their research outputs according to the same probability distribution, it is natural to assume that these authors are independent from one another due to lack of interaction. Equally reasonable is the assumption of mutual independence of the counts $\nu_j(k)$ for each individual author $k = 1, \dots, K$. In this notation, the pooled numbers M_j are given by

$$M_j = \sum_{k=1}^{K} \nu_j(k) \qquad (j = 0,1,2,\dots). \tag{5}$$

Then, according to the law of large numbers, for $K \gg 1$ we have

$$\frac{M_j}{K} = \frac{\nu_j(1) + \cdots + \nu_j(K)}{K} \approx E(\nu_j), \tag{6}$$

where E stands for expectation (statistical mean) and a random variable ν_j represents the counts $\nu_j(k)$. Recalling (4), the mean number of papers per author is approximated as

$$\frac{M}{K} = \sum_{j\geq 0} \frac{M_j}{K} \approx \sum_{j\geq 0} E(\nu_j) \qquad (j = 0,1,2,\dots). \tag{7}$$

provided that the series on the right is convergent. Thus, combining formulas (3), (6) and (7), we obtain the link between the power-law frequencies f_j and the expected counts $E(\nu_j)$,

$$f_j \approx \frac{M_j/K}{M/K} \approx \frac{E(\nu_j)}{\sum_{j\geq 0} E(\nu_j)} \qquad (j = 0,1,2,\dots). \tag{8}$$

In practice, power law is often fitted in the tail of the frequency distribution, that is, for $j \geq j_*$, with a suitably chosen truncation point j_*. This leads to readjustment of the normalizing constant c_a in the frequency formula (1). Fitting such a model to the data requires optimization over two parameters, a and j_*. Specifically, a natural heuristic tool to fit a truncated power-law model is by looking at the frequency plots (e.g., histograms) with logarithmic scales on both axes, whereby one seeks a straight-line fit, with the slope corresponding to $(-a)$ (cf. Nicholls, 1987). An alternative approach (Clauset, Shalizi & Newman, 2009), which provides the helpful smoothing of the discrete data, is via the complementary cumulative frequencies

$$F_j = \sum_{\ell \geq j} f_\ell \qquad (j \geq j_*). \tag{9}$$

Using again the log-log plots, a good fit corresponds to a straight line, with slope $(1 - a)$.

Generalized power-law model

Setting of the model

The generalized power-law (GPL) model introduced in this section is set out using two hyper-parameters n and m, interpreted as the mean numbers per author of citations and papers, respectively. These parameters can be estimated from the observed pooled corpus by

$$n \approx \frac{N}{K}, \qquad m \approx \frac{M}{K}. \tag{10}$$

There are also two shape parameters, $a > 2$ (akin to the power-law exponent) and $b > 0$. Namely, the citation frequencies are now assumed to be of the form (cf. (1))

$$f_j = \frac{Cm^{ab-1}}{(j + m^b)^a} \qquad (j = 0,1,2,\dots). \tag{11}$$

For small j ($j \ll m^b$) we have $f_j \approx C/m$, while for larger j we get a power-law dependence,

$$f_j \approx \frac{Cm^{ab-1}}{j^a} \qquad (j \gg m^b). \tag{12}$$

This may be viewed as an effective sewing of the formerly truncated lower values with the power-law tail. Recalling the link (8) between the pooled frequencies f_j and the individual expected counts $E(\nu_j)$, we have

$$E(\nu_j) \approx m\, f_j = \frac{C}{(jm^{-b} + 1)^a}\,. \tag{13}$$

Hence, we can calibrate the model using (10) to make it consistent with the hyper-parameters,

$$m \approx \sum_{j \geq 0} E(\nu_j) = C \sum_{j \geq 0} \frac{1}{(jm^{-b} + 1)^a} \approx Cm^b \int_0^\infty \frac{dx}{(x+1)^a} = \frac{Cm^b}{a-1}, \tag{14}$$

and similarly

$$n \approx \sum_{j \geq 0} j\, E(\nu_j) = C \sum_{j \geq 0} \frac{j}{(jm^{-b} + 1)^a} \approx Cm^{2b} \int_0^\infty \frac{x\, dx}{(x+1)^a} = \frac{Cm^{2b}}{(a-1)(a-2)}\,. \tag{15}$$

Considering the ratio n/m (i.e., the mean number of citations per paper), we get

$$\frac{n}{m} \approx \frac{m^b}{a-2}, \qquad C \approx m^{1-b}(a-1). \tag{16}$$

Hence, b is expressed in terms of the hyper-parameters and the parameter a,

$$b \approx \frac{\log n + \log(a-2)}{\log m} - 1\,. \tag{17}$$

The shape parameter a can be fitted to the data using either ordinary least squares or a suitable version of the maximum likelihood estimation (cf. Nicholls, 1987).

Limit shape

It is useful to represent the citation profile of an individual author by ranking their papers according to the citation scores (i.e., accumulated numbers of citations) $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$; for example, $\lambda_1 = \max\{\lambda_i\}$ is the score of the most cited paper. The citation profile is succinctly visualized by the *Young diagram* formed by (left- and bottom-aligned) row blocks with $\lambda_1, \lambda_2, \ldots$ unit square cells; its upper boundary is the graph of the step function

$$Y(x) = \sum_{j \geq x} \nu_j \qquad (x \geq 0), \tag{18}$$

where ν_j are the author's citation counts. In particular, $Y(0)$ is their total number of papers.

A useful insight into the structure of the citations data may be available by looking at the shape of suitably rescaled Young diagrams when both n and m are large (Vershik, 1996). Specifically, set

$$A = m^b, \qquad B = n/m^b, \tag{19}$$

and consider the rescaled (expected) shape

$$E\big(\tilde{Y}(x)\big) = \frac{1}{B} E\big(Y(Ax)\big) = \frac{1}{B} \sum_{j \geq Ax} E(\nu_j). \tag{20}$$

Substituting expressions (13) and approximating the sum by an integral like in (14) and (15), it is easy to show that there is a limit shape given by

$$E\big(\tilde{Y}(x)\big) \approx \varphi(x;a) = \frac{a-2}{(x+1)^{a-1}} \qquad (x \geq 0). \tag{21}$$

Data analysis

Goodness-of-fit

In this section, we fit the two models discussed above to real citation data, collected in January 2020 for a small population of 113 authors identified as those who had published at least one paper in the *Electronic Journal of Probability* in the first 10 issues (January–October) of volume 24 in 2019 (https://projecteuclid.org/euclid.ejp/1546571125) and who are also featured on Google Scholar (https://scholar.google.com). The total counts for this data set are $K = 113$ (authors), $N = 245{,}567$ (citations) and $M = 15{,}400$ (papers), including $M_0 = 6{,}472$ papers with zero citations. Noting that the observed frequency $f_0 = M_0/M = 0.42$ is quite high, it was decided to omit the value at $j = 0$ in the GPL model fitting (as seen in Figure 2, this value does appear to be an outlier). According to (10), the hyper-parameters of the GPL are given by $m = (M - M_0)/K = 79.01$ and $n = \ 2{,}173.16$. In turn, the corresponding estimates for the shape parameters are $a = 2.50$ and $b = 0.60$, found numerically using the *optim* command in R.

The power-law model was fitted using a suitable truncation as explained after equation (8); the fitted values, obtained using the *poweRlaw* package in R (https://cran.r-project.org/web/packages/poweRlaw), are $a = 2.32$ and $j_* = 48$. As we will see, the goodness-of-fit of the power-law model is excellent, but a high value of j_* is disappointing. Note that if we opted to ignore truncation and tried to fit a power-law model in the entire range, the fitted value of the exponent would change to $a = 2.42$.

First, let us report the results for the GPL model regarding the match of the theoretical limit shape $y = \varphi(x; a)$ specified in equation (21). In Figure 1, this limit shape is plotted in scaled-back coordinates for a better comparison with the data (represented by an empirical Young diagram as explained above), that is, $y = B\ \varphi(x/A; a)$, with the scaling coefficients given by formula (19) and estimated from the data as $A = 13.75$ and $B = 158.08$.

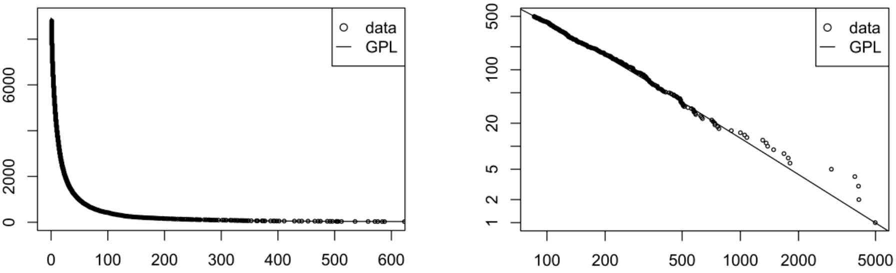


Figure 1. Observed data in the Young diagram representation and the fitted limit shape, with estimated parameters $a = 2.50$ and $b = 0.60$. The left panel illustrates a match for lower values of citations, while the right panel shows the tail comparison in the log-log coordinates.

We see from Figure 1 that the fit of the GPL is remarkably accurate, especially over a large initial part of the citation spectrum. To inspect details of the tail behavior, we use the log-log scale, revealing some minor discrepancies due to few extreme points.

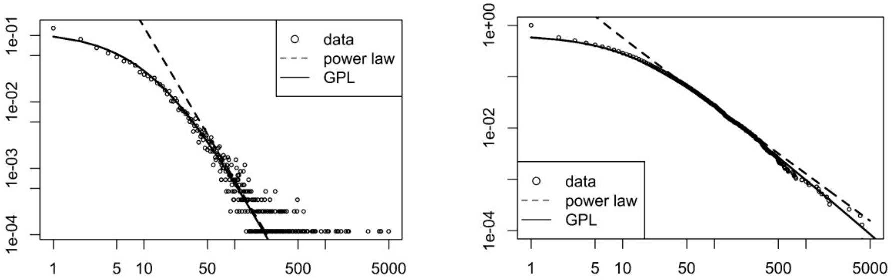


Figure 2. Log-log plots comparing data and the fitted models. The left panel shows frequencies f_j, while the right panel features complementary cumulative frequencies F_j. The left part of the dashed lines indicates an extrapolation of the power law below the truncation point $j_* = 48$. The observed frequency at $j = 0$ (zero citations) is an outlier.

Next, we compare the fit of both models to the data using standard frequency plots (Figure 2). Not surprisingly, the power law works extremely well at the tail but it is useless for smaller values of j. In contrast, the GPL strongly outperforms the power law over the initial range (despite a visible outlier at $j = 0$) but also works equally well at the tail.

Estimation of the h-index

Let us now look at what our models can tell about the h-index. According to Egghe & Rousseau (2006), in the non-truncated power-law model the h-index is estimated by the formula

$$h \approx m^{1/a}, \tag{22}$$

where $m = M/K$ is the mean number of papers per author. In our case, $m = 79.01$; using the non-truncated estimate $a = 2.42$ we get $h = 6.08$, while if we (formally) use the truncated fit $a = 2.31$ then the estimated value of h would slightly change to 6.63.

In the GPL model, the h-index geometrically corresponds to inscribing a biggest square inside the empirical Young diagram of citations. Using the limit shape (21) and scaling back using the coefficients A and B, we find that h is the approximate solution of the equation

$$h\left(\frac{h}{A}+1\right)^{a-1} = B(a-2). \tag{23}$$

Solving this equation numerically yields $h = 20.29$. Comparing with the empirical (mean) value $h = 17.52$, we see that the GPL estimation is superior to that of the power law, which cannot capture the true h-index lying deep below the truncation point $j_* = 48$.

Conclusion

In this paper, we have attempted to connect the pooled frequencies modeled via power law with individual citations per author. We have also introduced the generalized power-law (GPL) model which has been demonstrated to fit to the real data well in the entire range of citations, unlike the power law which frequently needs to get truncated at the beginning. In a real data set that we have studied, the GPL model reveals an inflated frequency of zero citations, which draws attention to a possible lack of impact in scientific output. An additional novel feature of the GPL model is that it possesses a deterministic limit shape which can be useful in estimating informative function of the citation data such as the h-index. Finally, it would be interesting to compare our GPL model with alternative fitting approaches (Sichel, 1985).

Acknowledgments

Work by Ruheyan Nuermaimaiti is part of her PhD research at the School of Mathematics, University of Leeds. The authors are grateful to Anna Senkevich for helpful insights concerning data retrieval from the web and to an anonymous referee for suggested bibliographic references.

References

Clauset, A., Shalizi, C.R. & Newman, M.E.J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51, 661–703.

Coile, R.C. (1977). Lotka's frequency distribution of scientific productivity. *Journal of the American Society for Information Science*, 28, 366–370.

Egghe, L. & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, 69, 121–129.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16569–16572.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16, 317–323.

Nicholls, P.T. (1987). Estimation of Zipf parameters. *Journal of the American Society for Information Science*, 38, 443–445.

Price, D.J.D.S. (1965). Networks of scientific papers. *Science*, 149, 510–515.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B – Condensed Matter and Complex Systems*, 4, 131–134.

Sichel, H.C. (1985). A bibliometric distribution which really works. *Journal of the American Society for Information Science*, 36, 314–321.

Vershik, A.M. (1996). Statistical mechanics of combinatorial partitions, and their limit shapes. *Functional Analysis and Its Applications*, 30, 90–105.

The Challenges of Building Publication Collections in the Wild: The Case of German Art History

Jenny Oltersdorf[1] and Jochen Gläser[2]

[1] *jenny.oltersdorf@tu-berlin.de*
TU Berlin, Hardenbergstr.16-18, 10623 Berlin (Germany)

[2] *jochen.glaeser@tu-berlin*
TU Berlin, Hardenbergstr.16-18, 10623 Berlin (Germany)

Abstract

We address a methodological problem that needs to be solved whenever databases of thematically or regionally circumscribed literature are built from scratch. How does one collect and delineate the scholarly literature produced by a national sub-community in a particular discipline? For our study of communication processes in the German Social Sciences and Humanities (SSH) we need a near-complete collection of the scholarly literature of German national sub-communities in a humanities field, namely art history. In this paper we discuss the definition of a German national sub-community of art history and the first step in the methodology of collecting its literature, namely the creation of a seed dataset.

Introduction

In this paper we report our solutions to a methodological problem that needs to be solved whenever databases of thematically or regionally circumscribed literature are built. How does one collect and delineate the scholarly literature produced by a national sub-community in a particular discipline? The common approach to creating specialized sets of publication metadata utilizes one of the two commercial databases Scopus or Web of Science and creates a subset of their data by thematic or regional delineation, for which sophisticated methods have been developed (Leydesdorff & Cozzens, 1993; Lewison, 1999; Aksnes et al., 2000; Zitt & Bassecoulard, 2006). This approach implicitly or explicitly assumes that the databases index all the relevant literature.

The premise of near-complete coverage has been shown to be wrong for the social sciences and humanities (as well as for many other disciplines, Moed, 2005: 119-136). This is why scholars build dedicated databases for the study of communication processes in these fields (Ardanuy et al., 2009; Colavizza & Romanello, 2017; Hammarfelt, 2012). The methodology of building these databases is rarely discussed. This is unfortunate for two reasons. First, the methodology and particularly the decisions on which publications to include affect the outcomes of studies for which the database is used. Insufficient documentation of methodological decisions thus unnecessarily limits the replicability of findings. Second, building such databases can be very labour intensive, and the exchange of methodologies could increase efficiency.

For our study of communication processes in the German Social Sciences and Humanities (SSH) we need a near-complete collection of the scholarly literature of German national sub-communities in one social science field (international relations) and one humanities field (art history) (Gläser & Oltersdorf, 2019). In this paper we discuss the definition of a German national sub-community of art history and the first step in the methodology of collecting its literature, namely the creation of a seed dataset.

Conceptual background: How to define German art history?

The collection of publication metadata reported in this paper is the prerequisite of an investigation of SSH communication processes with a focus on national communication substructures and the use of languages (Gläser & Oltersdorf, 2019). Collecting publication metadata of all German publications in art history in a specific time period requires a clarification of what we actually mean by “German art history”. The two conceptual problems that must be solved are identifying German *art history* and identifying *German* art history.

We start from the observation that in the sciences, social sciences and humanities knowledge is produced not by disciplines but by smaller research specialties (Chubin, 1976). Research specialties are scientific communities whose collective identity is based on members’ perception of collectively advancing a shared body of knowledge. Members observe this body of knowledge, derive problems and approaches to solving them from it, and offer solutions to these problems to other community members as contributions to the shared knowledge (Gläser, 2006). Membership is a perceived rather than ascribed status.

From this perspective art history as a discipline is a system of overlapping specialties, each of which develops a specific body of knowledge. The specialties share a general object and an historical approach. Members of an art history specialty perceive themselves as such and express this perception by using the community’s knowledge and offering contributions to the community. This means that community members become identifiable by their publications. Overlaps of communities result from researchers using knowledge of more than one community and addressing their contributions to more than one community,

The definition of membership as perceived rather than ascribed and the overlaps of communities create problems for the identification of community members that are relevant to our task. The overlaps do not end with the disciplinary boundaries of art history but extend into overlaps with, among others,

- communities investigating objects closely related to the arts (e.g., architecture),
- communities sharing the historical approach of art history communities (e.g., general history),
- communities utilizing interdisciplinary approaches to the study of art (e.g., area studies), and
- communities investigating contemporary art with non-historical approaches (e.g., art criticism, the sociology of arts).

As for all other communities, membership in an art history community can be partial (when researchers split their activities between several communities) and temporary (when researchers move between communities in the course of their career).

Most scientific communities are international in that they have members working in many countries. Some international scientific communities, particularly in the social sciences and humanities, develop regionally or nationally intensified communication processes in which contributions are primarily addressed to community members from the same region or country. These national sub-communities can develop for several reasons including the existence of a large number of community members in a country, a shared non-English language, or the existence of nationally or regionally specific research objects that are of lower interest to researchers outside the country or region (Gläser, 2006: 184-186).

Operationalization: Methodological decisions

Strategy

From the conceptual considerations follows that we are looking for a collection of publications that draw on the same knowledge (have references in common) and represent the formal communication process of the German sub-community (cite each other). Thus, the ultimate decision about the delineation of German art history as a nationally intensified discourse will be based on citation relations and shared references. The strategy of producing such a dataset is in accordance with the common approach in bibliometrics in that we create a seed dataset that is subsequently enhanced by adding publications to which it has citation relations and reduced by excluding publications without such relations. However, it differs in its first step due to the problems of creating the seed data set. In this section, we first discuss possible sources of publication metadata from which a seed data set could be created. We then describe the strategy for creating the seed data set and the methodological decisions made.

The most important consideration for creating the seed data set is that its composition must not introduce bias in the publication collection. To reduce the danger of bias, the seed data set needs to be large enough to cover the variation of publications, in particular with publication types and languages used.

Sources for the seed data set: Dead ends

Metadata of publications in German art history can be collected from a variety of sources, each of which has its specific limitations.

Curated collections that include references have insufficient coverage

The two main databases that index publications, Web of Science (WoS) and Scopus, have the main advantage that they include the references of indexed publications and links between publications that are indexed in the database and appear in reference lists. There is widespread consensus that they are not suitable for analyzing SSH communication processes due to their insufficient coverage of the SSH literature and their bias against particular publication types and publication languages (Engels et al., 2018; Hicks, 2004; Kulczycki et al., 2020). We included publications indexed in these databases but could not use them as exclusive source of a seed data set due to the limitations of their coverage of the literature.

Collections of links to internet sites do not provide publication metadata

The most prominent collection of links to internet sites that host publications, Google Scholar (GS), indexes web documents and links them with web documents that cite them or are referenced by them. GS differs from WoS and Scopus in its collection strategy, which is inclusive and automated. The Google crawlers are indexing any academic document that can be found on the web. The coverage of literature by GS thus depends on the internet publication strategies of authors and publishers. As a consequence, older publications are poorly covered as neither the publications nor the citations are (yet) online. Detailed information about the coverage of GS is not available. GS does not include publication metadata but points to internet sites that do. Publications are not thematically indexed.

For these reasons, GS can be used to extend the seed data set via citation relations but not to build it. Another source that links items and references is the member driven platform CrossRef. The coverage is determined by the metadata provided by the CrossRef community, and is therefore uneven and ill understood. Citation based analysis in CrossRef is limited to 59% of the items, for which free and open reference information are provided by publishers and cooperating institutions (Martín-Martín et al., 2021). Therefore, CrossRef can be seen as a broker of metadata between organizations (Hendricks et al., 2020).

No national publication data collections exist in Germany

Confronted by the limitations of existing databases, some countries have chosen to develop national research information systems, with complete bibliographic coverage of the scholarly output at research institutions e.g., Belgium (*VABB-SHW Database*) or Croatia (*CROSBI*) (Sīle et al., 2018). No such national-level research information system exists in Germany. There is not even a register of institutional bibliographies, as not all higher education institutes provide a publicly available bibliography of their researchers' output.[1]

Specialized international collections insufficiently cover the German literature

Répertoire international de la littérature de l'art (RILA) and *Bibliography of the History of Art* (BHA) are specialized art historical bibliographies. They include articles from over 1,200 journals, and cover material published between 1975 and 2007. The *International Bibliography of Art* (IBA) is the successor to BHA. Its description of coverage is only vague. Its website claims that it retains the editorial policies of BHA without providing further explanation.[2] Rinehart (1990) described the content and policy of BHA as follows: "The overall framework remains postclassical Western art. […] Practical considerations have naturally played a part, and the general feeling has been that what is done should be done well and comprehensively before undertaking more. […] The policy of BHA is to reflect, not to define, the changing boundaries of the discipline." (Rinehart, 1990: 134)

IBA includes several types of publications but is limited to 500 international journals, many of which are covered only partially. About 60% of the entries are from non-English language publications. Despite its apparent comprehensiveness IBA's coverage of German art history is very patchy. An author search for five randomly selected German art historians showed no results or returned only small proportion of their research output.

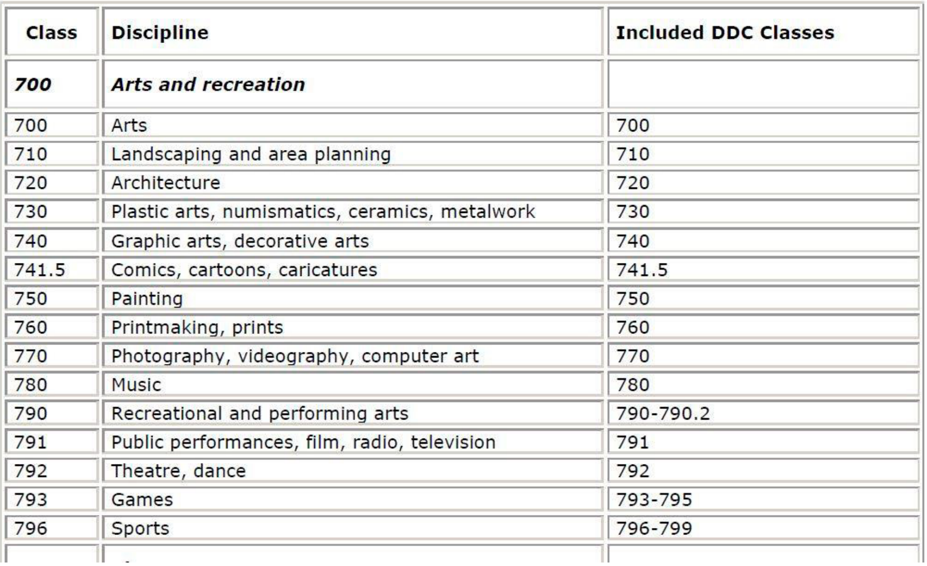
<u>German library collections are decentralized and difficult to collate</u>

The organization of library collections through classifications, subject headings or thesauri is context-dependent and subject to permanent revisions (Chen, 2013; Hjørland, 2013; Nohr, 1996). Creators of a classification impose a particular view of the knowledge on the users by organizing the research field. Any field of knowledge may, in fact, be classified from different epistemological perspectives which, in turn, may produce quite different classifications and publication collections (Hjørland, 1998). This is not problematic as long as publication metadata are collected by one library consistently using one system.[3] Unfortunately, the German library infrastructure is fragmented, with several libraries forming a distributed national library. Scientific libraries are organized into six library networks that rely on different classification schemas for indexing, without an integrating layer in form of a union catalogue of all six networks (Seefeldt, 2017). There are several research libraries that are specialized in art history, and belong to different networks.

A comparison of *Dewey Dezimalklassifikation* (DDC), *Basisklassifkation* (BK) and *Regensburger* Verbundklassifikation (RVK) reveals the complexity of that delineation problem. All three classifications are used for indexing in the German library system, and all three classifications consider art history differently. We illustrate the problem by comparing the classifications of art history and architecture because the distinction of art history from architecture is one of the main delineation issues. RVK includes a main class *art history* and assigns the subclass architecture to it. BK also includes a class *art history* as part of main *class Kunstwissenschaft (*art studies*)*, but subsumes architecture as part of the main class *Bauwesen* (civil engineering). DDC, in contrast, does not contain a main class with the name art history at all. Instead, it uses a class *Arts and recreation* (Figure 1). This is not only a question of labelling. Rather, the creators of DDC did not intend a classification of the discipline of art history, but actually wanted to classify *the arts*. This inevitably leads to a completely different classification, in which there is no place for a class *art history*.

This fragmented structure with different classification schemas for indexing, unlike meta data standards, divergent depth of indexing, and the lack of a union catalogue hinder retrieval and download of meta data for delineation.

Each of the sources we reviewed represents only a subset of existing publications. That leads to the assumption that completeness in coverage can only be considered regarding predefined indicators. As all databases deal with a certain degree of publication loss we decided to turn the focus from the question of coverage to the question of how to best identify the community.



Class	Discipline	Included DDC Classes
700	***Arts and recreation***	
700	Arts	700
710	Landscaping and area planning	710
720	Architecture	720
730	Plastic arts, numismatics, ceramics, metalwork	730
740	Graphic arts, decorative arts	740
741.5	Comics, cartoons, caricatures	741.5
750	Painting	750
760	Printmaking, prints	760
770	Photography, videography, computer art	770
780	Music	780
790	Recreational and performing arts	790-790.2
791	Public performances, film, radio, television	791
792	Theatre, dance	792
793	Games	793-795
796	Sports	796-799

Figure 1: DDC Subject categories class 700

Constructing a seed data set: methodological decisions

In the light of the problems with sources of publication metadata, we decided to start from the self-identification of researchers as art historians and to extract publication metadata from their publication lists, and reference information from additional sources or full texts of publications (Oltersdorf et al., 2020). This approach required two methodological decisions. First, researchers in art history must be distinguished from those in adjacent disciplines. Secondly, researchers belonging to German art history must be distinguished from those belonging to other national art history communities.

Who is an *art historian*?

Self-descriptions of art history are not free of contradictions and regional variation, and disciplinary boundaries have been moved in several directions by the art history community itself. We decided to accept a broad definition of art history that we found in the review literature (Belting, 2008). This definition is organized along three indicators: time, object of investigation and territory. In terms of time, art history is limited to the Christian-influenced eras starting in the 4th century to the present. With regard to the present, art history must be distinguished from art criticism as there are fundamental differences in the methods applied (Michalski, 2015). Regarding the territory, art history is conceived as a discipline that deals with research objects generated and shaped by the European conception of art. This includes the art of North and South America as well as the Christian cultures of the Near East and the Caucasus. Research objects of art history include architecture as well as sculpture, painting, graphic arts, craftwork, up to photography and new media.

We operationalized this delineation by identifying organizational structures of universities that label themselves as being art historical according to that definition. We then identified all affiliated researchers and parsed the publication lists from the universities' websites. It is quite clear that the organizational structures do not always coincide with the cognitive classification of publications (Bourke & Butler 1998). It is also undeniable that the completeness of the literature lists presented at the universities' websites varies. Despite these limitations, the publications thus identified are very likely to include the majority of publications for the time frame considered. Since they also include all publication types that are common in art history, they form a suitable seed data set that can be completed in subsequent steps.

Which publications belong to *German* art history?

The decision to parse literature lists from researchers affiliated to a German university with a department or institution that corresponds to the definition of art history also implies a decision of who belongs to German art history. From the definition of a national sub-community, several indicators for membership in the German art history sub-community could be derived. A first publication-based indicator to distinguish *German* art history from other countries could be the language. However, German is used as a publication language by a variety of researchers. In addition to researchers from Austria or Switzerland, other researchers also publish in German because art history as a field is multilingual. By the same token, using the publication language as an indicator would exclude the large number of publications written in languages other than German, which is not acceptable in a multilingual discipline as German art history (Gläser & Oltersdorf, 2019).

An author-based approach to delineating German art history could further consider indicators like membership in a national professional association or institutional affiliation. Membership of a particular professional association, is a fuzzy indicator in so far as membership of such an association is not coextensive with community membership. Depending on the respective society's rules, students, retired researchers who are no longer active or researchers from adjacent fields can become members. It is also unlikely that all active researchers are members of the professional association.

Considering this theoretical implications and limitations for operationalization, we decided to accept only the institutional affiliation to a research unit that self-describes as conducting research in art history as an indicator of community membership. Therefore, we included all researchers affiliated to the selected university institutes regardless of their biographical background. We used this information to build the seed data.

This selection does not consider some specifics of German art history. First, three non-university research institutions abroad (listed in Table 1), and the *Zentralinstitut für Kunstgeschichte*, a non-university research institution in Munich, are part of the German art history network but are not included in our search strategy for the seed data set.

Table 1: Non-university research institutions abroad

Institute	*Location*	*Funded by*
Kunsthistorisches Institut in Florenz	Florence (Italy)	Max-Planck-Society
German Center for Art History	Paris (France)	Max Weber Foundation
Bibliotheca Hertziana – Max Planck Institute for Art History	Rome (Italy)	Max-Planck-Society

This observation reveals two delineation problems. First, it turns out that the German art history subcommunity may extend beyond German borders. Secondly, art historians also work in organisations other than universities. As for the first problem, it is quite likely that art historians who work outside Germany are members of a German art history subcommunity. However, even for those working in foreign subsidiaries of German non-university research associations, community membership would first have to be established. After all, it is participation in a specific discourse that makes a researcher a member of the German art history subcommunity. Since there is no clear causal link between German funding for a research institute and the employed researchers' membership in a German national research subcommunity, we excluded publications of these researchers from our seed dataset.

The decision is less easy for non-university organisations in Germany because important research is contributed by art historians working in museums and art galleries (Dilly, 2008). The German *Institute for Museum Research* identified 718 art museums in 2018 (Institut für

Museumsforschung, 2019). They exhibit artefacts from art and architecture, arts and crafts, ceramics, church treasures and ecclesiastical art, film to photography. Information about art historians affiliated with museums is not publicly available. These community members can only be identified through their publications. However, most of these publications are edited volumes published by the institution, which also features as editor. Therefore, identifying the edited volumes does not help. Since publication metadata for book chapters are rarely available, author information could only be obtained from (printed) full texts. We therefore decided not to include these publications in the seed data set.

Discussion

By presenting our considerations on the task of identification and definition of a regional research community we highlighted the complex interplay between conceptualization, operationalization, and practical opportunities. The decisions made for the construction of the seed data set for our database are specific to the German context. For example, we had to respond to severe shortcomings of the German library infrastructure. Approaches that utilize library collections might be more successful in countries with a more centralized or more standardized library system. While some of our decisions may have been made differently, it is difficult to see how a seed dataset that covers most German art history publications in a specified time frame could be made.

Conclusions

The need to manually build a database of publication metadata highlights how many decisions that have consequences for research are left to commercial database providers whenever their data are used. Three points are of particular importance. First, even for fields that are well covered by the databases coverage is not complete. Bibliometrics should establish a standard for discussing coverage when it is likely to affect results, similar to the discussion of non-responses to surveys.

Secondly, our discussion of who should count as a German art historian highlights possible consequences of delineations based on author addresses. In many cases, delineation by address is the correct approach. For example, if effects of institutions are investigated, researchers working in a country are the population affected by its science policy institutions (with the interesting exception of research organisations abroad). However, if national sub-communities or discourses are to be investigated, attention should be paid to the specific consequences of different decisions on whom to include.

Thirdly, the internet offers the opportunity to link existing metadata collections rather than creating data collections from scratch. The number of tools providing this service is increasing. On the one hand, these services are important because they save resources by facilitating access to data. On the other hand, using these tools is not without danger because the data quality depends on third parties – the many providers of metadata - and is therefore more difficult to manage. In addition, curation in the sense of delineation is often fraught with problems as the selection criteria of the data providers vary substantially or are not known at all.

As a general conclusion, we would like to raise the question whether we know enough about the way in which the data in commercial databases are curated. Bibliometricians have pointed out for a long time bibliometrics' unusual situation as a research community not in control of its data. At the same time, we base our reputation as researchers on the analysis of these data. Until attempts to create community-controlled alternative data sources, we should be wary.

Acknowledgments

This research was supported by the German Federal Ministry of Education and Research (Grant 01PU17022).

References

Aksnes, D. W., T. B. Olsen and P. O. Seglen (2000). "Validation of bibliometric indicators in the field of microbiology: A Norwegian case study." *Scientometrics* 49(1): 7-22.

Ardanuy, J., Urbano, C., & Quintana, L. (2009). A citation analysis of Catalan literary studies (1974–2003): Towards a bibliometrics of humanities studies in minority languages. *Scientometrics*, 81(2), 347-366. https://doi.org/10.1007/s11192-008-2143-3

Belting, H. (Hrsg.). (2008). *Kunstgeschichte: Eine Einführung* (7., überarb. und erw. Aufl). Reimer.

Bourke, P. and L. Butler (1998). "Institutions and the map of science: matching university departments and fields of research." *Research Policy* 26(6): 711-718.

Chen, K. (2013). Dynamic Subject Numbers Replace Traditional Classification Numbers. *Knowledge Organization*, 40(3), 60–168.

Chubin, D. E. (1976). "The Conceptualization of Scientific Specialties." *Sociological Quarterly* 17(4): 448-476.

Colavizza, G., & Romanello, M. (2017). Annotated References in the Historiography on Venice: 19th–21th centuries. *Journal of Open Humanities Data*, *3*(2). http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.9/

Dilly, H. (2008). Einleitung. In *Kunstgeschichte Eine Einleitung*. Dietrich-Reimer- Verlag.

Engels, T. C. E., Istenič Starčič, A., Kulczycki, E., Pölönen, J., & Sivertsen, G. (2018). Are book publications disappearing from scholarly communication in the social sciences and humanities? *Aslib Journal of Information Management*, *70*(6), 592–607.

Gläser, J. (2006). *Wissenschaftliche Produktionsgemeinschaften. Die soziale Ordnung der Forschung*. Frankfurt a. M., Campus.

Gläser, J., & Oltersdorf, J. (2019). Persistent Problems for a Bibliometrics of Social Sciences and Humanities and How to Overcome Them. In *Proceedings of the 17th International Conference on Scientometrics & Informetrics (ISSI-2019) (2-5 September 2019, Rome, Italy)* (Vol 1, p. 1056–1067). http://www.issi-society.org/proceedings/issi_2019/ISSI%202019%20-%20Proceedings%20VOLUME%20II.pdf

Hammarfelt, B. (2012). Harvesting footnotes in a rural field: Citation patterns in Swedish literary studies. *Journal of Documentation*, 68(4), 536–558. https://doi.org/10.1108/00220411211239101

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1), 414–427. https://doi.org/10.1162/qss_a_00022

Hicks, D. (2004). The Four Literatures of Social Science. In *Handbook of Quantitative Science and Technology Research* (p. 473–496). Kluwer Academic Publishers.

Hjørland, B. (1998). The Classification of Psychology: A Case study in the Classification of a Knowledge Field. *Knowledge Organization*, 24(4), 162–201.

Hjørland, B. (2013). Theories of Knowledge Organization—Theories of Knowledge. *Knowledge Organization*, 40(3), 169–182.

Institut für Museumsforschung. (2019). *Statistische Gesamterhebung an den Museen der Bundesrepublik Deutschland für das Jahr 2018* (Nr. 73).

Kulczycki, E., Guns, R., Pölönen, J., Engels, T. C. E., Rozkosz, E. A., Zuccala, A. A., Bruun, K., Eskola, O., Starčič, A. I., Petr, M., & Sivertsen, G. (2020). Multilingual publishing in the social sciences and humanities: A seven-country European study. *Journal of the Association for Information Science and Technology*, 71(11), 1371–1385.

Lewison, G. (1999). "The definition and calibration of biomedical subfields." *Scientometrics* 46(3): 529-537.

Leydesdorff, L. and S. E. Cozzens (1993). "The Delineation of Specialities in Terms of Journals Using the Dynamic Journal Set of the SCI." *Scientometrics* 26(1): 135-156.

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1), 871–906.

Michalski, S. (2015). *Einführung in die Kunstgeschichte*. Wissenschaftliche Buchgesellschaft.

Moed, H. F., Ed. (2005). *Citation Analysis in Research Evaluation*. Dordrecht, Springer.

Nohr, H. (1996). *Systematische Erschließung in deutschen Öffentlichen Bibliotheken (Classificatory Subject Analysis in German Public Libraries)*. Harrassowitz.

Oltersdorf, J., Mironenko, A., & Gläser, J. (2020). A Workflow for Creating Publication Databases from Scratch. In *Lockdown Bibliometrics: Papers not submitted to the STI Conference 2020 in Aarhus* (p. 37–50). https://www.sos.tu-berlin.de/menue/discussion_paper/

Rinehart, M. (1990). BHA/Bibliography of the history of art/Bibliographie d'histoire de l'art. *Art Documentation: Journal of the Art Libraries Society of North America*, 9(3), 134–136.

Seefeldt, J. (2017). *Verbundsysteme in Deutschland Entstehung der Bibliotheksverbünde*. https://bibliotheksportal.de/informationen/bibliothekslandschaft/bibliotheksverbuende/

Sīle, L., Guns, R., Sivertsen, G., & Engels, T. C. E. (2017). *European databases and repositories for Social Sciences and Humanities research output*. ECOOM & ENRESSH. https://doi.org/10.6084/m9.figshare.5172322

Sīle, L., Pölönen, J., Sivertsen, G., Guns, R., Engels, T. C. E., Arefiev, P., Dušková, M., Faurbæk, L., Holl, A., Kulczycki, E., Macan, B., Nelhans, G., Petr, M., Pisk, M., Soós, S., Stojanovski, J., Stone, A., Šušol, J., & Teitelbaum, R. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. *Research Evaluation*, 27(4), 310–322.

Zitt, M. and E. Bassecoulard (2006). "Delineating complex scientific fields by a hybrid lexical-citation method: An application to nanosciences." *Information Processing & Management* 42(6), 1513-1531.

[1] Nor is there a European publication database for SSH research output, mainly because institutional and national data sources use different data models as well as different data collection and validation procedures (Sīle et al., 2017).

[2] https://www.proquest.com/iba/index (accessed: 29.01.2021)

[3] For example, the Dutch national library comprehensively collects all Dutch publications and publications concerning the Netherlands.

A Typology of Research Discovery Tools: Bibliometric Redundancy vs. Bibliometric Variety

Andreas Pacher[1]

[1] *andreas.pacher@tuwien.ac.at*
TU Wien, Bibliothek, Resselgasse 4, A-1040 Vienna (Austria)
and Vienna School of International Studies, Favoritenstraße 15, A-1040 Vienna (Austria)

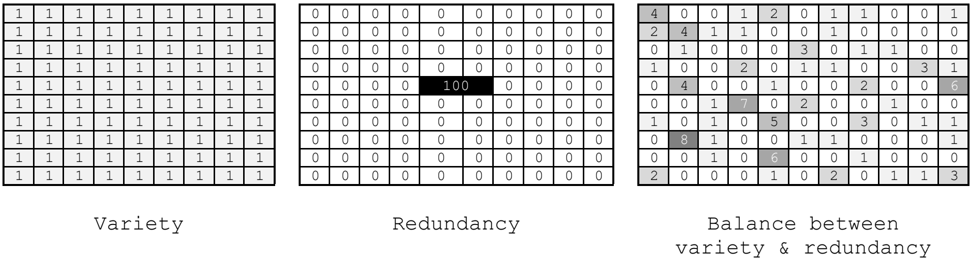
Abstract

Bibliometric links between scientific publications may exhibit redundancy (i.e. expectable co-occurrences between publications) or variety (i.e. unique and original reference patterns). The roots of aggregate redundant or variety-prone citation interlinkages may lie in the way how researchers find relevant publications. But how can we make sense of the potential landscape of all such research discovery tools? This paper conceptualizes a typology based on the common variable of redundancy/variety. On the redundancy-reproducing end of the spectrum are machines that draw from co-citations or keyword queries (such as academic search engines and paper recommender systems), while the variety-end harbours services that enable categorial browsing or that suggest publications randomly (such as journals' tables of contents or random paper bots on Twitter). Currently, redundancy-reproducing research discovery tools proliferate. In contrast, there is little awareness about variety-generating ones. But if the scientific system demands a balance between both redundancy and variety to bring forth innovations, then the designs and discussions of future research discovery tools should better appreciate aspects of variety.

Introduction

Imagine a scientific system in which every research finding must be cited once, and only once. Such a condition would ensure maximum *variety* of reference patterns within the system. One may also imagine the opposite, namely a structure under which only a single publication must be cited in every work, while all other research papers remain uncited. This pattern would amount to a repetitive reproduction of bibliometric *redundancy*.

Neither scenario is wished for; both push the scientific system into entropy. The first scenario (of maximum variety) does so by constantly bringing forth surprising citations without a predictable structure conducive to an accumulative knowledge growth. The second one (of maximum redundancy) engenders entropy due to a permanent repetition of the same. A well-functioning scientific system, in contrast, would combine both variety and redundancy (cf. the right illustration in Figure 1): A new publication conveys familiarity by referencing already-known works, but it simultaneously energizes the system with new information requiring new citations – e.g. through an innovative perspective, an unusual theory, an original question, or novel data.



1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1

Variety

0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	100	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

Redundancy

4	0	0	1	2	0	1	1	0	0	1
2	4	1	1	0	0	1	0	0	0	0
0	1	0	0	0	3	0	1	1	0	0
1	0	0	2	0	1	1	0	0	3	1
0	4	0	0	1	0	0	2	0	0	6
0	0	1	7	0	2	0	0	1	0	0
1	0	1	0	5	0	0	3	0	1	1
0	8	1	0	0	1	1	0	0	0	1
0	0	1	0	6	0	0	1	0	0	0
2	0	0	0	1	0	2	0	1	1	3

Balance between
variety & redundancy

Figure 2. Three variants of how 100 citations may be distributed among 100 (in the center: 99) scientific publications. – Left: Variety (each paper accrues 1 citation). – Center: Redundancy (each paper is uncited, with the exception of a single publication which accrues 100 citations). – Right: Some papers are cited more than others in a seemingly random distribution.

The same observation applies to research discovery tools, i.e. to machines that draw from a set of scientific publications so as to recommend a selection of them to users. Research discovery would be inadequate if it recommended publications merely once and then never again (maximum variety), or if it always recommended the very same paper (maximum redundancy). Like scientific reference patterns, the landscape of research discovery tools should instead find a balance between variety and redundancy.

While neither the maximum-variety nor the maximum-redundancy-scheme seems realistic, the current structure of scientific publications does approximate the latter version. The standard assumption departs from a Pareto distribution whereby 80 % of all citations go to just 20 % of all research outputs (Nisonger, 2008). Thus, the modal publication remains little-cited, while a few 'highly cited' papers accrue most attention. This redundancy-prone pattern finds its reflexion in the designs and discussions of many research discovery tools. Rather than enabling variety, they reproduce bibliometric redundancy through query-based and citation-based algorithms. On a closer look, even the literature on 'serendipity', i.e. on the accidental discovery of useful publications for research purposes, falls in this vein; its paradoxical aim of transforming the 'irrelevant' and 'useless' into the 'relevant' and 'useful' leaves variety in a blind spot. However, the scientific system requires a quick tempo of internal variety to bring forth innovations; it is through non-redundant combinations of research findings that original discoveries and conceptual leaps take place (Acemoglu et al., 2016). The focus on redundancy should thus be accompanied by a clearer incorporation of bibliometric variety into the current discussions and designs of research discovery.

This paper addresses this blind spot by conceptualizing a typology of possible research discovery tools based not on the useful/useless-distinction, but on the conceptual dimensions of bibliometric redundancy/variety. In brief, it proposes a schematic typology ranging from devices that draw from co-citations and semantic queries (which are both at the 'redundancy'-spectrum) to machines that recommend publications based on certain categories and randomness (which are located at the 'variety'-end). On the one side are therefore platforms that detect bibliographic couplings, academic search engines and paper recommenders that systemically reproduce bibliometric redundancy. On the other side are category-based browsing-featured tools (e.g. institutional repositories, a set of journal table of contents, or a list of latest or trending papers) and instruments based on sheer randomness (such as random paper bots). The closer a tool reaches the variety-end of the spectrum, the more its recommendations are unbiased from publication dates, measurements of impact, research fields, or extant citational links and academic or semantic networks.

This paper elucidates the common variable behind all research discovery tools and suggests a more conscious combination of both redundancy- and variety-based platforms in order to address the scientific system's structural demand for plurality. It also departs from earlier attempts to categorize research discovery services by providing a universal variable common to all such platforms. Previous attempts, in contrast, confined themselves to a narrower focus, such as to query-based variants (e.g. whether a tool enables Boolean search terms or not; cf. Kim & Rebholz-Schuhmann, 2008), on personalized and non-personalized services (Kotkov et al., 2016), or on paradoxical distinctions such as 'useless' versus 'useful' publications (e.g. Ge et al., 2010).

The following section illuminates the theoretical concepts of informational redundancy and variety. The paper then briefly reviews the current state of research discovery tools and the literature on serendipity, finding that they all tend to reproduce scientific redundancy. To address this one-sidedness, a conceptual typology of research discovery tools based on the distinction of redundancy and variety follows that review. A final section discusses the advantages and limitations of this typology and concludes with suggestions regarding the discoveries further research on research discovery may await.

Variety and Redundancy

Every information provides a surprise: For it momentarily appears and disappears, resolves uncertainty, and renders any given context slightly different than before (Bateson, 1999; Luhmann, 1992, p. 391). However, the surprisefulness inherent to an information can vary by degrees. For instance, there is less surprise in cases of informational *redundancy*, that is, when the knowledge of one information allows inferences about other (yet unknown) information (Shannon, 1951). The revelation of the other (redundant) datum would then seem expected, predicted and superfluous. A non-redundant system, on the other hand, harbours *variety*, or a rich plurality of distinct events so that an observer's knowledge of one information does not disclose another one (Luhmann, 1992, p. 438). At their extreme, redundancy reproduces self-circularity in a permanent repetition of the same, while variety increases a system's openness towards more and more stimuli.

Applying this distinction of redundancy and variety to the issue of interest here, namely to a bibliometric 'information encountering' regarding scholarly publications (Erdelez, 1999), redundancy would be akin to co-citation patterns: With bibliometric redundancy, a given scientific reference increases the probability of the presence of another specifiable reference. For instance, say that *AAA* is often co-cited alongside *BBB* as both study similar aspects of *X*. Now, if a paper on *X* cites *AAA*, it will most likely cite *BBB*, too. Or, if a research discovery tool recommends *AAA* based on a specific user-input, it will not be a large surprise if it also recommends *BBB* based on the same user-input. Only specific works profit from this redundancy, while all others remain in an uncited and thus invisibilized spot. Another paper on *X*, namely *CCC*, may have 0 citations so far despite its topical relevance to both *AAA* and *BBB*. *AAA* and *BBB* may nevertheless accrue dozens of co-occurring citations while *CCC* is left with none. In total, there may be 100 papers on that topic of *X*, but only the most probable two of them were realized: the "maximum information content" of the system remains in a non-realized state (Leydesdorff et al., 2018, p. 1182). It is thus that bibliometric redundancy gets reproduced.

Redundancy and variety may serve as two opposing sides of a distinction, but they are not antipoles: The scientific system can increase both its redundancy and its variety simultaneously. A paper can cite both *AAA* and *BBB* on the one hand and *CCC* on the other hand. What the system requires is a mixture of both, conveying familiarity so as to enable the temperate growth of a given knowledge-field, but, at the same, rendering the familiar more interesting (Davis, 1971) by injecting a surprising novelty into it, a new perspective, a new theory, a new datum, with new scholarly references.

Note that the distinction variety/redundancy does not refer merely to the relation between two publications, but rather to an aggregate pattern that grows over time from the viewpoint of a specific topic *X* (see Table 1 for an illustration). Thus, if a new paper on the topic of *X* cites *AAA* and *BBB*, it reproduces redundancy (because these references were structurally predictable given that these two are the most cited papers on the topic *X*), but if it cites *CCC* or *DDD*, it generates variety (for these two references are relatively surprising within the topic of *X* as they are little-cited within that field). The concepts of variety and redundancy are thus not properties of publication-dyads, but rather of a whole network of multiple publications.

Neither redundancy nor variety bear an inherent value of 'good' or 'bad' (Luhmann, 1992, p. 436). Redundancy is not *per se* to be condemned, but rather exercises vital functions in communicative settings. If there was no redundancy, i.e. if every paper was cited or recommended only once, one would require greater labour in searching for a specific reference or publication, rendering an informational loss more probable. Redundancy thus saves time and serves as an insurance against research waste. Great variety, on the other hand, is not *per se* to long for, but it rather generates an unpredictable system whose capacity may not suffice to carry the immense plurality of unique information within it. As publications would quickly fall into oblivion, bibliometric variety would also risk a high amount of wasted research. Both one-sided redundancy and one-sided variety ultimately render the system entropic: Maximum redundancy would mean entropy due to an almost complete loss of any energizing surprises; maximum variety would mean entropy due to an almost complete loss of predictable structures.

Table 1. Three examples of how a citation can generate bibliometric redundancy or variety.

Cited paper	*Property before it is cited*	*When a new paper on topic X cites it, this leads to…*
Paper I	Already well-cited in the field of *X*	Redundancy (because it reproduces extant citation pattern)
Paper II	A little-cited paper	Variety (because it opens up a novel reference)
Paper III	Well-cited on the topic of *Y*, but has never been cited on the topic of *X*	Variety (because it bridges two topics in an original way)

These abstract-theoretical propositions find empirical illustrations in various scientometric studies. For example, redundant references serve as orientations for scholars to determine whether a given paper matches their fields and interests (Riviera, 2013); they enable an iterative assessment and re-assessment of studies; by allowing citation-couplings, they ensure that efforts of a study conducted at one moment in time may bear fruit over decades (van Raan, 2004); and, quantified into citation counts and similar measures, they aid in the numerical and comparative observation of career trajectories among researchers, concepts, and publication outlets (Garfield, 1972; Leydesdorff, 1998; Min et al., 2018).

The tendency to reproduce redundancy can, however, engender detrimental effects. A bibliometrically redundant system lacks the influx of new information necessary to energize scientific innovations; it instead tends to reproduce the same contents over and over. Such predictable patterns then become “not … very useful, or very interesting” (Maloney & Conrad, 2016, p. 9), reducing “exposure to less popular (and perhaps more interesting) information” (Maloney & Conrad, 2016, p. 4). What is more, outdated works may grow in citation counts while new publications are overlooked (Merton, 1968), leading to a perceived injustice against scholarly newcomers. The number of uncited or low-cited works may then be perceived as problematically reflecting wasted research efforts and public funds (van Leeuwen & Moed, 2005). And as soon as redundant citation patterns can be measured on a large scale, they are subject to ‘cheating’ behaviors on the levels of researchers, institutions and journals – with the symptomatic accidents of high self-citation rates, a culture of ‘publish or perish’, predatory journals, irreproducible research, and an inflation in publication and citation counts.

Bibliometric variety in citation patterns may serve as a remedy against these perils of systemic redundancy. Original citation patterns, for instance, offer a robust indicator for innovative research; for it is the “cross-pollination” of ideas across disciplines (Erdelez, 1999, p. 4) that pushes the evolution of knowledge further (Bornmann et al., 2019; Wang et al., 2017). Through surprising references, variety “stimulates creativity by illustrating new connections, connections that were not […] anticipated” by the structure of citational co-occurrences (Race, 2011, p. 140). It extends a given research field's capacity for processing information and thus corrects the narrowness of its former bibliometric structure (Luhmann, 1992, p. 450). In addition, it helps integrating societally marginalized researchers into the scientific system.

However, variety alone would likewise exhibit perilous effects. With variety modelled to the extreme, researchers would lose important markers for orientation within hitherto familiar research fields, which discourages a cumulative knowledge growth illustrated in the self-reproducing interconnectivity of citation patterns (Riviera, 2013). If translated into a growing number of references per publication, variety may lead to 'superficial' citations (van Wesel et al., 2014) and, ultimately, to an inflation of indicators that are supposed to measure scientific impact, resulting in the necessity for ever-more complicated algorithms to measure how influential a scientific publication was (Mingers & Leydesdorff, 2015). As such measurements deeply affect researchers', funders' and students' behaviour – such as through world university rankings (Dearden et al., 2019) – artificially skewed data due to an overgrown bibliometric variety could multiply decision errors across various sectors of life.

A balance of both redundancy and variety is thus functionally important in ensuring the autopoiesis of the scientific system. But do research discovery tools and the discussions thereof meet the systemic demand for balancing both aspects? The next section argues that there may be an overemphasis on redundancy what to the detriment of the equally vital aspect of variety.

Research Discovery and Serendipity: An Overemphasis on Redundancy

Research discovery tools require a pre-selected input to narrow down the range of recommendable publications. Such a pre-selection is necessary to avoid total randomness; for instance, a research discovery tool should not recommend a nocturnal poem from Ancient Japan to a user interested in discussions of how contemporary Western politicians co-manage global crises in the Middle East. At the minimum, the genre of recommended documents should be confined to scientific publications. A mere genre-level pre-selection, however, is still too broad to be useful. The user interested in the political dimension of Middle East conflicts in the beginning of the 21st century should, for instance, not be recommended papers on gene editing or on chemical engineering simply because they are scientific publications. The pre-selection could rather be narrowed down to a specific research domain (e.g. political science), or even narrower, to detailed keywords (e.g. 'Political Conflicts', 'Middle East', 'Sovereignty'). Whatever the degree of narrowness, the pre-selection ensures that the recommended paper does not deviate too much from the user's needs.

For a discovery tool to be useful to researchers, the input-output-scheme should not be overly trivial (von Foerster, 1960). The literature on informational relevance and serendipity stresses how algorithms should go beyond a narrow semantic distance between the user-input and the output in the form of a recommend publication (Race, 2011). However, instead of using the concepts of redundancy and variety, they depart from the distinction of irrelevant/relevant, accidental/planned or useless/useful. They principally ask how irrelevant 'accidents' can be made relevant and 'purposeful' so as to render the 'useless' 'useful' (Race & Makri, 2016), to 'expect the unexpected' (Maloney & Conrad, 2016, p. 1), or to 'search for the unsearchable' (Campos & Figueiredo, 2001, p. 159). In the end, reflections on serendipity paradoxically equal *x* to its contrasting *y* (Spencer Brown, 1969), finding that the distinct are identical to themselves as they propose the 'useless' to be undistinguishable from the 'useful' (McCay-Peet & Toms, 2015, p. 1464).

A consequence of the paradox is that while the literature on serendipity aims at producing 'novel' and 'surprising' paper recommendations, i.e. at enhancing bibliometric variety, they ultimately offer technical suggestions that remain in a redundancy-reproducing realm. Recent surveys of serendipity in research recommender systems, for instance, explicitly leave 'non-personalized' approaches in the blind environment while confining their reviews to ‘personalized algorithms’ (Kotkov et al., 2016, p. 180). These, however, are *per definitione* oriented towards a target user’s revealed preferences (cf. Abbas & Niu, 2019; Bai et al., 2019). It is the researcher’s pre-existing catalogue of scientific publications (or their semantic proxies) that serves as the departure point for all these allegedly serendipitous recommender tools. Whatever the path they take in order to reach the output of a paper recommendation, all the tools orientate themselves towards the input of their user’s past behaviour, citational links, and existing networks – even if they use algorithms harnessing machine learning (Rakhshani et al., 2020), topological structures of academic communities or heterogeneous information networks (Boussaadi et al., 2020; Ma & Wang, 2019), a re-ordered ranking of ‘similarity scores’ (Afridi et al., 2020, p. 226) or complex co-citation analyses (Sakib et al., 2020). Thus, be they content-based or utilize collaborative filtering, be they graph-oriented or use hybrid methods – all these services first necessitate a user-input and then design an algorithmically distanced, but nevertheless unescapably proximate association to that user’s research interests. If the paper recommendation is a function of the user-input, then the tools cannot avoid the output’s proximity to the user. As every output is ultimately based on existing bibliographic or semantic co-occurrences (though with some algorithmic distancing), from a systemic viewpoint, all these so-called serendipitous research discovery tools are tendentially repetitive in terms of fostering a systems-wide bibliometric redundancy – despite their pledge to enhance variety.

The distinction between the useless and the useful seems too paradoxical to be useful in itself; for whatever the algorithmic design of a retrieval system, it will always and unavoidably exhibit a bias emphasising selected weighting functions based on an input. The best tool is then not a non-biased, but merely a ‘least biased’ model (Wilkie & Azzopardi, 2014). Instead of using the elusive concepts of the serendipity-literature (McBirnie, 2008), as in useful/useless papers or accidental/intentional discovery, this paper prefers to build a conceptual approach based on the distinction of redundancy/variety to differentiate research discovery tools.

Both the dominant landscape of research discovery tools and the related literature on serendipity focus on the reproduction of *redundancy* while leaving out, or only retaining a paradoxical vision towards, *variety*. But how would variety-centered research discovery tools look like? A conceptual typology based on the dimensions of redundancy and variety may aid in clarifying how the variable common behind all research discovery services can be operationalized at both ends

A Typology

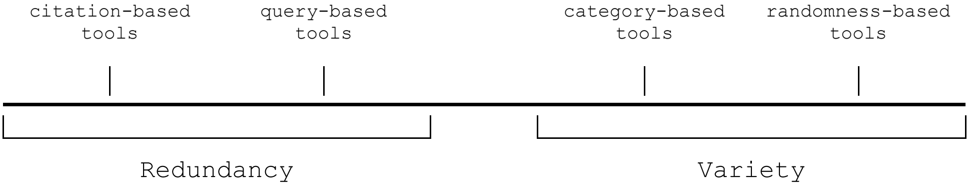


Figure 2. A typology of research discovery tools ranging from those that reproduce bibliometric redundancy (comprising citation-based and query-based tools) to those that generate bibliometric variety (comprising category-based and randomness-based tools).

The present section suggests that all research discovery tools can be located within a dimension ranging from (bibliometric) redundancy to (bibliometric) variety. At the redundancy end of the spectrum are tools that at least indirectly reproduce extant reference patterns between scientific publications. They may be citation-based (citational proximity) or query-based (semantic proximity). Other tools are located at the variety end of the spectrum. These variety-generating ones may be category-based (e.g. finding publications in a common journal or by a common researcher or in a common scientific domain) or randomness-based. Figure 2 illustrates the typology.

Redundancy-Generating Tools

First, citation-based research discovery tools are most prone to generating bibliometric redundancy for it is their very function to reproduce extant reference patterns. They indulge in "the gravity of highly cited papers" (Maloney & Conrad, 2016, p. 4). The user-input comes in form of a token for a scientific publication (or a set of multiple publications), such as by having the user insert a publication's title or DOI. The tool then looks for the citations to or from that publication; it may also detect second-order references, i.e. references to publications that cite or are cited by those publications which are citationally interlinked to the user input. Examples for citation-based tools are *CitationGecko*, *ConnectedPapers* or *CoCites*. Many query-based search engines likewise contain functions to reproduce reference patterns, such as, most prominently, *Google Scholar*'s feature to list citations to a certain publication.

A second type of research discovery tools at the redundancy-end of the spectrum are query-based ones. Many of these are search engines that require keywords to search for papers containing similar strings. *Google Scholar* is primarily built around this function. With varying functionalities (e.g. advanced search functionalities, Boolean operators, semantic web-searches), other prominent databases likewise work in this fashion: *Microsoft Academic*, *Dimensions*, *Web of Science*, *Scopus*, *Semantic Scholar, LENS* etc. mainly revolve around a search bar into which users type in their queries. Some recommender systems may likewise draw from semantic distance-formulae to narrow down their paper recommendations to users. In addition, tag-based platforms that enable users to attribute categorical labels to scientific publications in a manner of a 'collaborative filtering' operate on the same principle. Such tools (like *CiteULike*) aggregate semantic commonalities across research outputs (Heck et al., 2011). Whatever the precise approach, query-based tools are similar to citation-based platforms in that they are prone to reproduce bibliometric redundancy; semantic similarities thus serve as a proxy for extant inter-citation linkages.

User profile-based recommender systems likewise fall within this redundancy-end of the spectrum; they may either follow the logic of co-citations or of semantic proximities. Their main difference is that their user-input does not come in form of an actively inserted information of publications, but rather in the form of algorithmically generated user profiles. For instance, *Mendeley Suggest* or *ResarchGate*'s recommender system look at the papers stored by a user to recommend proximate publications. As it is their very task to generate user-near suggestions, they cannot but reproduce bibliometric redundancy.

Variety-Generating Tools

We have seen how citational proximity and semantic similarity reproduce citational redundancy across the scientific system. On the other side are tools that enhance variety. But how do these variety-generating research discovery tools operate?

First, the discovery tool may be based on the selection of certain categories. Such category-based platforms often facilitate a 'creative browsing' through publication lists (Bawden et al., 2011). There is a multitude of such instruments, but they are seldom analyzed as full-fledged research discovery tools, even though they serve the same function of conveying links to scientific publications to interested users. Even a *curriculum vitae* (CV) of an individual scholar can be subsumed under this category: It enables a profile-based discovery of research. Institutional repositories operate on the same basis; they allow for an institution-based research discovery. A more common way to explore publications uses a journal-based approach, where one browses through a specific outlet's table of contents, either in print or digitally (an experience that can be enhanced through RSS feeds or e-mail alerts). Some systems aggregate tables of contents from a greater set of journals, such as *Current Contents* and *JournalToCs*. If that set of journals is defined by a specific research domain, we have discipline-based platforms such as those narrowed down to Criminology (e.g. *Criminology Papers*) or Philosophy (e.g. *Philosophy Paperboy*) – these tools then regularly send out titles of the latest papers published in a specific research domain to their users.

Most of these category-based platforms simply look at the *latest* papers based on their publication dates. Others may incorporate additional indicators, such as whether the publications of a specific research domain published in a certain date range (e.g. past 30 days) are 'trending'. For instance, the *Observatory of International Research* (*OOIR*) not only lists the latest papers from political science and related domains, but also ranks them according to their *Altmetric Attention Scores* based on social media mentions and news attention.

Thus, discovery tools of this sort list the latest or trending papers according to a pre-selected category, whereby that chosen unit may be defined by a certain scholar, by a shared research institution, by a common journal or a set of journals, or by a research domain. The more general the categorial unit, the less redundant the research discovery tool: The knowledge of the category does not reveal knowledge of specific publications anymore, as opposed to co-citation patterns. Most importantly, the listed outputs of these category-based research discovery tools are detached from extant reference networks. Such tools do not directly reproduce bibliometric redundancy, but enhance variety instead.

However, even category-based platforms are not completely free from biases that may be conducive to an indirect redundancy. They at least implicitly demand a more or less specific category (e.g. an individual scholar, a set of journals, or a certain scientific discipline) to which the user may already have developed systemic affinity. In addition, browsing-featured tools usually list the *latest* papers – but if bibliometric patterns bear a biased citation life cycle towards newer publications to the detriment of older ones (Cano & Lind, 2005), then even these platforms reproduce this tendency and thereby serve to nourish at least a small degree of systemic redundancy. Tools that list *trending* papers are perhaps likewise stained by a structural bias whereby papers in highly-cited 'top journals' will more likely be 'trending' in social media and news platforms (cf. Ferreira Araujo, 2020). Thus, even category-based platforms may indirectly contribute to citational redundancy.

In contrast, an even more variety-enhancing research discovery platform would list new *and* old papers, trending *and* non-trending papers, discipline-specific *and* discipline-deviating papers with equal probability. Such truly variety-enhancing tools would be one based on a large sample of scientific publications and on a high degree of randomness (Maloney & Conrad, 2016, p. 4); they may not even exist yet. A fictitious example would be a huge corpus of the scientific literature in all research domains covering various centuries that would regularly (e.g. hourly) spit out a random publication of this corpus through, say, a Twitter bot. It may 'recommend' a philosophical paper on epistemology from a German journal of the 1820s, and at the next instance one medical publication from a low-ranked journal from the 2000s, then a highly-cited sociological work from the 1970s. The pattern of outputs should be unpredictable, with every new recommendation equally improbable as other ones, gushing out a truly 'blind variation' (Campbell, 1960) that would constantly nourish the evolution of science. Only such a tool would ensure a high degree of liberty from extant biases, thus serving as a research discovery platform that is veritably variety-enhancing for the whole system of science. And it would then be up to the system's evolution – or, as earlier centuries would say, up to individual geniuses and their sagacity – to fetch these random surprises and to de-randomize them by transforming them into scientific knowledge (Luhmann, 1992, pp. 466–467).

Discussion

This paper proposed a conceptual typology of all possible research discovery tools based on the dimensions of bibliometric redundancy (the reproduction of extant citational links between scientific publications) and variety (the generation of novel ties between research outputs). The exemplary tools harboured by this typology comprise a highly heterogeneous landscape, including a journal's table of contents, a researcher's CV, scholarly databases, academic search engines, institutional repositories, or Twitter bots that regularly recommend random scientific works to its users.

This typology demands universality, in that it makes visible a common variable to be imputed behind *all* research discovery tools. It thus departs from previous attempts to categorize scientific discovery services in a narrower manner, e.g. by typifying merely query-based approaches (e.g. whether they allow 'boolean queries' or not; cf. Kim & Rebholz-Schuhmann, 2008, p. 454). Furthermore, this paper indicates how the extant and oft-used research discovery tools are designed such that they suppress the element of systemic surprise. The discussions and designs of research discovery instead tend to reproduce citational redundancy. The ideal and extreme type of a variety-enhancing tool does not even seem to exist yet: There is not yet the fictitious bot that spits out random scientific publications from a huge corpus from centuries of publications of which each work has an equal probability to be listed, regardless of the publication's age, discipline, authors, citational impact, or publication outlet. This absence may be telling of how much meta-scientific reflections and implementations have so far focused on redundancy while leaving the issue of variety in a blind spot.

In effect, if researchers only used *Google Scholar* and *Mendeley Suggest* to discover relevant publications, they would mainly reproduce bibliometric redundancy; this "too precise" discovery "can end up reinforcing habits rather than exposing students and researchers to new information, sharply limiting the researcher's view of the world" of scientific papers (Maloney & Conrad, 2016, p. 9). They would only discover publications whose contents do not deviate too much from a semantic or co-citational closeness to the pre-specified user-input (e.g. a typed-in query of keywords, or a list of publications one has already read). To use the example from above, if the users have already read or are interested in the topic of *AAA*, then they would be recommended *BBB* because of their dense co-occurrences – that is perfectly well-expected, but not innovative. Finding only papers of topical proximity, such tools would minimize the element of surprise and thereby inhibit innovative cross-pollinations from distant fields. The scientific system would only contain a slow tempo of variety, with only rare energization through new information and original reference patterns.

Systemic bibliometric citation patterns begin with the way researchers find publications to skim, read, interpret, evaluate and cite (cf. Crestani et al., 1998). The algorithms of research discovery tools and information retrieval systems thus trickle down to aggregated reference patterns within the scientific system. Their design may thereby generate greater or lesser system-wide entropy. It is for this reason that more awareness of the shared variable on the basis of which a landscape of various such tools can be mapped, and greater efforts at balancing the opposing sides of the dimensions at the level of research discovery would be wished for. Variety-enhancing tools can be utilized even with a "vague, fuzzy, or even unspoken information need, when users do not quite know what they're looking for" (Maloney & Conrad, 2016, p. 2). Without this appreciation, researchers will "discount serendipity because it is not viewed as a formal search strategy" (Race, 2011, p. 140). Extant platforms (for instance, *Altmetric*) could be harnessed more consciously for variety-enhancing purposes (Holbrook, 2019, p. 6); this typology just serves as a first step in conceptually illuminating such potentials.

The distinction of redundancy/variety also shows that there is no obvious "paradox of control" (McBirnie, 2008, p. 611) behind serendipity; it is not inherently rare and does not need "both luck and skill" and is not "both unpredictable and yet can be cultivated" (Copeland, 2019, p.

2386). It is, instead, an empirically observable, systemic distribution of references to scientific publications that have some of its roots in the machines that recommend them to researchers.

Despite all practical implications from this paper, there is, of course, "no single web service that satisfies all demands" (Kim & Rebholz-Schuhmann, 2008, p. 452). Science needs both redundancy (for orientating oneself through familiar structures) and variety (for generating novel and innovative knowledge). Indeed, research discovery platforms often do contain multiple services built within them, some of them redundancy-reproducing, others of them variety-enhancing: A journal's website may show both the latest papers (variety-enhancing) and its most-cited ones (redundancy-enhancing). To make it more complicated, and to repeat a statement from above, redundancy and variety are not contrasts; both can be heightened at the same time. *Google Metrics* (which is built within *Google Scholar*) may serve as an example: It enables a category-based browsing tool (i.e. variety-enhancing function) whereby each publication list is ordered by a journal-level h-index, and thus by citational impact (i.e. redundancy-enhancing function).

Innovation does not always originate from specific approaches to research discovery, but also through external triggers. Any cue in life can offer potential ideas – be it a look outside the window, be it a Twitter acquaintance, be it a lab meeting over a coffee. However, as they all remain outside a conceptualization of science as an autopoietic system that reproduces itself through instances of verity-claiming publications (Luhmann, 1992), they are environmental 'irritations' that do not reflect the aggregate structure of research links *within* the system itself, but only the venues through which the science takes inspiration from *outside*. Otherwise the whole world could be deemed a 'research discovery tool', which might be too much of a semantic overstretch.

How to proceed further in the research of research discovery tools? For scientific explorations of research discovery tools, one may generate a clearer empirical mapping of extant discovery services based on the dimensions proposed here. One may also devise experiments of exposing researchers only to specific research discovery tools for a limited period of time, and see how it affects their citation patterns. As regards technical implementations, one may generate a truly variety-enhancing tool as that fictitious random paper bot outlined above, and see how it affects its users' serendipitous experiences and their drive for innovation. Whatever the next directions for scientific research and technical implementations of research discovery tools – there is much variety to discover.

References

Abbas, F., & Niu, X. (2019). Computational Serendipitous Recommender System Frameworks: A Literature Survey. *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, 1–8.

Acemoglu, D., Akcigit, U., & Kerr, W. R. (2016). Innovation network. *Proceedings of the National Academy of Sciences*, 113(41), 11483–11488.

Afridi, A. H., Yasar, A., & Shakshuki, E. M. (2020). Facilitating research through serendipity of recommendations. *Journal of Ambient Intelligence and Humanized Computing*, 11(6), 2263–2275.

Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., & Xia, F. (2019). Scientific Paper Recommendation: A Survey. *IEEE Access*, 7, 9324–9339.

Bateson, G. (1999). *Steps to an Ecology of Mind*. University of Chicago Press.

Bawden, D., Foster, A., & Rafferty, P. (2011). Encountering on the road to Serendip? Browsing in new information environments. In *Innovations in Information Retrieval* (pp. 1–22). Facet Publishing.

Bornmann, L., Tekles, A., Zhang, H. H., & Ye, F. Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. *Journal of Informetrics*, 13(4), 100979.

Boussaadi, S., Aliane, H., Abdeldjalil, O., Houari, D., & Djoumagh, M. (2020). Recommender systems based on detection community in academic social network. *2020 International Multi-Conference on Organization of Knowledge and Advanced Technologies*, 1–7.

Campbell, D. T. (1960). Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review*, 67(6), 380–400.

Campos, J., & Figueiredo, A. D. de. (2001). Searching the Unsearchable: Inducing Serendipitous Insights. *Proceedings of the Workshop Program at the Fourth International Conference on Case-Based Reasoning, ICCBR 2001*, 159–164.

Cano, V., & Lind, N. (2005). Citation life cycles of ten citation classics. *Scientometrics*, 22(2), 297–312.

Copeland, S. (2019). On serendipity in science: Discovery at the intersection of chance and wisdom. *Synthese*, 196(6), 2385–2406.

Crestani, F., Lalmas, M., Rijsbergen, C. J. V., & Campbell, I. (1998). “Is This Document Relevant? . . . Probably”: A Survey of Probabilistic Models in Information Retrieval. *ACM Computing Surveys*, 30(4), 25.

Davis, M. S. (1971). That’s Interesting!: Towards a Phenomenology of Sociology and a Sociology of Phenomenology. *Philosophy of the Social Sciences*, 1(2), 309–344.

Dearden, J. A., Grewal, R., & Lilien, G. L. (2019). Strategic Manipulation of University Rankings, the Prestige Effect, and Student University Choice. *Journal of Marketing Research*, 56(4), 691–707.

Erdelez, S. (1999). Information Encountering: It’s More Than Just Bumping into Information. *Bulletin of the American Society for Information Science and Technology*, 25(3), 26–29.

Ferreira Araujo, R. (2020). Communities of attention networks: Introducing qualitative and conversational perspectives for altmetrics. *Scientometrics*, 124(3), 1793–1809.

Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation. *Science*, 178(4060), 471–479.

Ge, M., Delgado-Battenfeld, C., & Jannach, D. (2010). Beyond accuracy: Evaluating recommender systems by coverage and serendipity. *Proceedings of the Fourth ACM Conference on Recommender Systems*, 257–260.

Heck, T., Peters, I., & Stock, W. G. (2011). Testing Collaborative Filtering against Co-Citation Analysis and Bibliographic Coupling for Academic Author Recommendation. *Proceedings of the 3rd ACM RecSys’ 11 Workshop on Recommender Systems and the Social Web*.

Holbrook, J. B. (2019). Designing responsible research and innovation to encourage serendipity could enhance the broader societal impacts of research. *Journal of Responsible Innovation*, 6(1), 84–90.

Kim, J.-J., & Rebholz-Schuhmann, D. (2008). Categorization of services for seeking information in biomedical literature: A typology for improvement of practice. *Briefings in Bioinformatics*, 9(6), 452–465.

Kotkov, D., Wang, S., & Veijalainen, J. (2016). A survey of serendipity in recommender systems. *Knowledge-Based Systems*, 111, 180–192.

Leydesdorff, L. (1998). Theories of citation? *Scientometrics*, 43(1), 5–25.

Leydesdorff, L., Johnson, M. W., & Ivanova, I. (2018). Toward a calculus of redundancy: Signification, codification, and anticipation in cultural evolution. *Journal of the Association for Information Science and Technology*, 69(10), 1181–1192.

Luhmann, N. (1992). *Die Wissenschaft der Gesellschaft*. Suhrkamp.

Ma, X., & Wang, R. (2019). Personalized Scientific Paper Recommendation Based on Heterogeneous Graph Representation. *IEEE Access*, 7, 79887–79894.

Maloney, A., & Conrad, L. (2016). *Expecting the Unexpected: Serendipity, Discovery, and the Scholarly Research Process*. SAGE Publications Inc.

McBirnie, A. (2008). Seeking serendipity: The paradox of control. *Aslib Proceedings*, 60(6), 600–618.

McCay-Peet, L., & Toms, E. G. (2015). Investigating serendipity: How it unfolds and what may influence it. *Journal of the Association for Information Science and Technology*, 66(7), 1463–1476.

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, 159(3810), 56–63.

Min, C., Ding, Y., Li, J., Bu, Y., Pei, L., & Sun, J. (2018). Innovation or imitation: The diffusion of citations. *Journal of the Association for Information Science and Technology*, 69(10), 1271–1282.

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, 246(1), 1–19.

Nisonger, T. E. (2008). The "80/20 Rule" and Core Journals. *The Serials Librarian*, 55(1–2), 62–84.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A Manifesto*.

Race, T. M. (2011). Resource Discovery Tools: Supporting Serendipity. In M. Pagliero Popp & D. Dallis (Eds.), *Planning and Implementing Resource Discovery Tools in Academic Libraries* (pp. 139–152). IGI Global.

Race, T. M., & Makri, S. (2016). Introducing Serendipity. In T. M. Race & S. Makri (Eds.), *Accidental Information Discovery* (pp. 1–13). Chandos Publishing.

Rakhshani, H., Latard, B., Brévilliers, M., Weber, J., Lepagnot, J., Forestier, G., Hassenforder, M., & Idoumghar, L. (2020). Automated Machine Learning for Information Retrieval in Scientific Articles. *2020 IEEE Congress on Evolutionary Computation (CEC)*, 1–7.

Riviera, E. (2013). Scientific communities as autopoietic systems: The reproductive function of citations. *Journal of the American Society for Information Science and Technology*, 64(7), 1442–1453.

Sakib, N., Ahmad, R. B., & Haruna, K. (2020). A Collaborative Approach Toward Scientific Paper Recommendation Using Citation Context. *IEEE Access*, 8, 51246–51255.

Shannon, C. E. (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal*, 30(1), 50–64.

Spencer Brown, G. (1969). *Laws of Form*. Allen & Unwin.

van Leeuwen, T. N., & Moed, H. F. (2005). Characteristics of journal impact factors: The effects of uncitedness and citation distribution on the understanding of journal impact factors. *Scientometrics*, 63(2), 357–371.

van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics*, 59(3), 467–472.

van Wesel, M., Wyatt, S., & ten Haaf, J. (2014). What a difference a colon makes: How superficial factors influence subsequent citation. *Scientometrics*, 98(3), 1601–1615.

von Foerster, H. (1960). On Self-Organizing Systems and Their Environments. In M. C. Yovits & S. Cameron (Eds.), *Self-Organizing Systems* (pp. 31–50). Pergamon Press.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8), 1416–1436.

Wilkie, C., & Azzopardi, L. (2014). Best and Fairest: An Empirical Analysis of Retrieval System Bias. In M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky, & K. Hofmann (Eds.), *Advances in Information Retrieval* (pp. 13–25). Springer International Publishing.

World research and European news-media reports of cancer risk factors: implications for public health communication

Elena Pallari[1] and Grant Lewison[2]

[1] *elena.pallari@outlook.com*
Health Services Research Center, Agiou Sergiou 33, Strovolos, 2037, Nicosia (Cyprus); University College London, MRC Clinical Trials and Methodology Unit, 90 High Holborn, London, WC1V 6LJ (UK)

[2] *grantlewison@aol.co.uk*
King's College London, Department of Cancer Policy, Guy's Hospital, Great Maze Pond, London SE1 9RT (UK)

Abstract

Cancer is now one of the leading components of the global burden of disease. Causes of cancer that are amenable to intervention are multiple: tobacco control closely followed by obesity treatment, including promotion of a healthy diet and physical exercise, remain the global priorities. We interrogated the Web of Science (WoS) from 2001 to 2020 to determine the numbers of papers describing research into 14 different possible risk factors causing cancer. These ranged in relative importance from tobacco to the consumption of excessively hot drinks (linked to oesophageal cancer), pollution (linked to lung cancer particularly) and also non-interventional genetic risks; how they had varied in time, and between different continental regions. Because many of these factors are subject to human behavioural choices, we also investigated how such research was being presented to the European public through newspaper reportage. About half of the factors that influence cancer incidence can be attributed to particular causes. They are led by tobacco use, but this is slowly declining in most high-income settings. Research outputs on some of these different factors in the continental regions correlated positively with their influence on the cancer burden. However, the selection of European newspaper stories was biased towards those risk factors that could be considered as being under the control of their readers. The multifactorial causes of cancers vary by country, but within-country differences are also observed in the research output. Reports of research in the mass media may have a role in the control of cancer.

Introduction

As public health improves worldwide, life expectancy increases, and there is demographic ageing. This means that cancer, and other non-communicable diseases (NCDs), have become ever larger contributors to the global disease burden, see Figure 1.

As with other NCDs, prevention is cheaper and more effective than cure, so there is an increasing effort world-wide to understand context-specific causes of cancer and to intervene (Ginsberg *et al.*, 2009; Zimmermann *et al.*, 2017; Hurry *et al.,* 2020). Many of these cancer risk factors are in theory under our control, such as our behaviour, our diet, and some aspects of our environment, but others are not, such as our genetics. In this study, we sought to determine which of these causative factors were being researched, and by which countries. We also sought to compare the global research agenda with the distribution of the different factors, see Table 1, and with how such research is being presented to the European public through the mass media. This is because a large part of the causal risks of cancer is contributed by activities that we choose to undertake such as whether we smoke, what and how much we eat, and where we live and work.

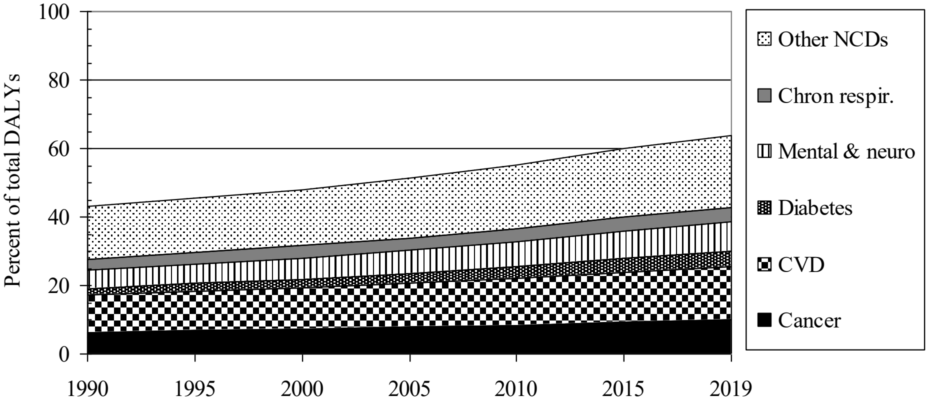


Figure 1. Contributors of different non-communicable diseases (NCDs; World Health Organization group 2) to the total global disease burden, measured in Disability-Adjusted Life Years (DALYs), 1995-2019. ***CVD = cardiovascular diseases. Data from the Institute for Health Metrics and Evaluation (IHME), University of Washington, Seattle.***

Table 1. Main causes of cancer risk, 1990-2019. Data from the Institute for Health Metrics and Evaluation (IHME), University of Washington, Seattle.

Cause		*1990*	*2000*	*2010*	*2019*	*Trend*
All		39.8	40.6	41.5	41.8	+
Behavioural		35.2	35.5	35.4	34.9	
of which:	Tobacco	24.5	24.5	24.3	23.6	–
	Dietary	5.69	5.63	5.67	5.55	
	Alcohol	4.83	4.98	5.07	5.17	+
	Sex, low act., drugs	4.75	4.78	4.70	4.68	
	High BMI	2.72	3.28	3.86	4.45	+
Air pollution		3.49	3.49	3.71	3.56	
Occupational		2.75	2.68	2.85	2.77	

Some of these causative factors are also constrained by government regulation. This is seen most obviously in the way that the tobacco industry has been circumscribed over recent decades, with restrictions on where people may smoke, the price of cigarettes, and controls on packaging (Flor *et al.*, 2021). One consequence is that smoking has, almost uniquely, decreased as a cause of cancer over the last three decades in many (but not all) settings, when others have increased, notably high Body Mass Index (BMI), that is obesity, and low physical activity.

Although there is an abundance of research on the epidemiology of cancer, there are remarkably few bibliometrics papers that have examined the outputs quantitatively, as opposed to systematic reviews and meta-analyses. The only one that we could find in relation to the topic was published 14 years ago (Ugolini *et al.*, 2007), and was reviewed later that year (Boffetta, 2007). There the matter rested until November of last year (Lin *et al*, 2020) when a team from Xiamen University examined Chinese publications, but only ones in the Surveillance, Epidemiology, and End Results (SEER) database, so their coverage was inevitably very incomplete.

In this study, we examined the world output of papers on cancer research that involved epidemiology as applied to cancer risk factors, and classified them by 14 subject areas, some of them corresponding to the modifiable cancer risk factors listed in Table 1. Most of the papers concerned an increase in the risks, but some described means of lowering it, such as taking more exercise or eating a good diet (*e.g.,* one rich in fruit and vegetables). These papers were classified by year, and by the continents of their authors.

We also analysed the research papers behind stories in 31 European newspapers reporting the epidemiology of cancer risk factors. These were all classified by inspection of the reported story. We classified the stories by the countries of the citing newspapers rather than those of their researchers.

Methodology

The papers about the risks of cancer were taken from the Web of Science (WoS, © Clarivate Analytics) from the 20 years, 2001-20. They were limited to articles and reviews, but with no language restriction. First, we selected all papers on cancer research by means of a complex filter based on 323 title words and 185 specialist cancer journals, which had been developed for other projects (see Begum *et al.*, 2018). It had a precision, p, of 0.95 and a recall, r, of 0.98, so it over-estimated the numbers of cancer papers by 3%. We then applied an epidemiology filter, consisting of 25 title words, such as:

*associat**, *cohort-study, determinant, exposure, incidence, predict**, *risk*

and a set of 35 epidemiology journals. They were selected from the list of all the journals used for cancer research where the papers had the selected title words, and whose names contained the strings *EPIDEM* or *PREVENTION* or *RISK*. This second filter was calibrated with reference to the outputs of researchers in epidemiology departments (Lewison, 1999), and had a precision of 0.78 and a recall of 0.89, so it over-estimated the numbers of epidemiology papers by 12%. The application of this second filter identified a true total of 126,594 papers, or 6.4% of the total in cancer research. The actual number of epidemiology papers identified in the WoS was 144,597, and these formed the database and were the ones whose characteristics were determined.

In order to select the main causes, we used the "compare" tab tool on the IHME website to identify the ones that were responsible for many cancer DALYs (Disability-Adjusted Life Years) or could be clearly defined by means of title words in their papers. [The settings were as follows: Display = Risk; Location = Global; Year = 2010; Age = All ages; Sex = Both; Level = 2.] We then inspected a large number of the cancer epidemiology papers and made a provisional list of the distinctive title words of the papers that could be used to identify those on each selected cause. This process was iterative, as the lists of title words were inevitably initially incomplete and was repeated until we created a comprehensive list of the title words to be used, some of which are reported in Table 4, below. We proceeded to download the long list of journals that had been used to publish the epidemiology papers. Journals whose names indicated that they were in one of the 14 selected subject areas (see Table 2) were then listed and used to select a second set of papers. We examined those papers that were so identified, but which were *not* identified by the provisional list of title words. This suggested both title words that could be added to the provisional list, and also that some of these specialist journals were not appropriate because many of the papers that they generated were not relevant. [They were often about the increased risks of *other* diseases because the patients already had cancer. However, we retained papers where the risk of cancer was increased because the patients were suffering from other diseases.]

Table 2. List of the 14 causes used for classification of the cancer epidemiology papers.

Code	*Subject area*	*Code*	*Subject area*
ALCO	Alcohol use	INFE	Infections
DENT	Dental health and	LIFE	Lifestyle choices
DIET	Food and drink consumed	OBES	Obesity and its *sequelae*
DRUG	Medication use	POLL	Pollution, and occupation
ECON	Socio-economic factors	RADI	Radiation, incl, medical use
GENE	Genetics	TOBA	Tobacco, incl. passive
HEAT	Hot drinks	UVRA	Ultraviolet radiation, tanning

The list of title words was then finalised for each subject area, complemented by the specialist journals, and applied to the epidemiology file in order to generate sets of papers, both in the world and in eight continental regions, see Table 3. [The geographical analysis presented here was by continent rather than by individual country in order to reduce the amount of data and make the results more easily understandable.]

Table 3. List of eight continental regions with their definitions

Region	*Short form*	*Definition*
Africa		
Asia		Including Taiwan and Turkey but excluding China
China		
Eastern Europe	E Eur	Including Russia and Ukraine
Latin America and the Caribbean	LatAmC	Including Mexico
North America	N Amer	Bermuda, Canada, USA, but excluding Mexico
Oceania		Australia, New Zealand and Pacific Islands
Western Europe	W Eur	The 15 European Union Member States in 2000, plus Iceland Norway, Switzerland and small nations

Table 4 lists the title words for five of the selected subject areas. They were designed to be used with the WoS software, and the hyphen indicates that the words must be next to each other in the titles. The WoS automatically includes plurals as well as singular forms, *e.g.*, *vitamins* as well as *vitamin* but wild cards (*) are needed to cover different forms such as *infected, infection, infects*. Some papers would have been classed in more than one of the 14 causes, and others not in any of them.

A description of the methodology used to record and analyse the newspaper stories about cancer research was given earlier (Pallari & Lewison, 2019). The study period was shorter, from 2002 to 2013, and stories were collected from selected newspapers in 20 European Union Member States (MS), 13 in "Western" Europe (the MS as of 1995, less Greece and Luxembourg), and seven in "Eastern" Europe who joined in the 21st century. Most of these stories were from the UK, Belgium, and Ireland, so they were not representative of the totality of European newspapers. The spreadsheet of stories contained their details, and also the bibliographic data of the research papers that they cited, taken from the WoS. These data were used to identify the cancer epidemiology papers, which were relatively numerous (991 out of a total of 3873 cited by the cancer research stories, or 26%). Their titles were individually inspected and the papers were marked with the tetragraph codes listed in Table 2.

Table 4. List of five of the 14 subject areas (causes) chosen for analysis of the cancer epidemiology papers, with the title words used to identify them.

Diet	*Genetics*	*Infection*	*Lifestyle*	*Obesity*
carbohydrate	allele	AIDS	exercise	bariatric surgery
consume	BRCA	e-coli	fitness	BMI
consumption	chromosom*	Escherichia	interval-training	body-fat*
dairy	deletion	Helicobacter	lifestyle	body-mass-index
deficiency	familial	HIV	phone	body-size
diet*	gene	infect*	physical activity	body-weight
food*	genetic	papillomavirus	running	Lorcaserin
fruit	hereditary	papilloma-virus	screen-time	obes*
grain	inherited	vaccin*	sedentary	overweight
meat	loci	virus	sexual-activ*	silhouette
nutrient	locus		sexual-behav*	waist
nutrition	mutation		sex-with-men	weight gain
vegetable	nucleotide		shift-work	
vitamin	polygenic		sitting	
	polymorphism		sports-activity	
	repeat		telephone	
	triple-negative		walking	
	variant		weight-training	

Results

Figure 1 shows the world outputs of cancer research papers from 2001 to 2020 (left scale), and the percentage that is represented by epidemiology (right scale). The former has increased by a factor of nearly four. This is partly real, but is also caused by the greatly increased journal coverage of cancer research in the WoS, from 2468 in 2001 to 6653 in 2020. Epidemiology as a percentage of cancer research increased slowly from 7.4% to 8.9% in 2014, and then dropped in 2015 as a result of the addition to the WoS of the Emerging Sciences Citation Index, and then continued to decrease to 6.9% in 2020. Of the countries with at least 1000 cancer research papers, the five Scandinavian ones (Iceland (29%), Finland (17%), Denmark and Norway (16%) and Sweden (15%)) published relatively the most on epidemiology. Those relatively the least active were Egypt (5.8%) and Switzerland and Turkey (6.2%).

Figure 2 shows the numbers of papers in each of the 14 subject areas, plotted on a logarithmic scale. There is a difference of almost three orders of magnitude between the largest (genetics) and the smallest (hot drinks). The latter seems to be locally or regionally important (mainly for oesophageal cancer) but does not feature among the main causes in Table 1. Table 5 shows the percentages of epidemiology papers in each of the 14 subject areas for the eight continental regions.

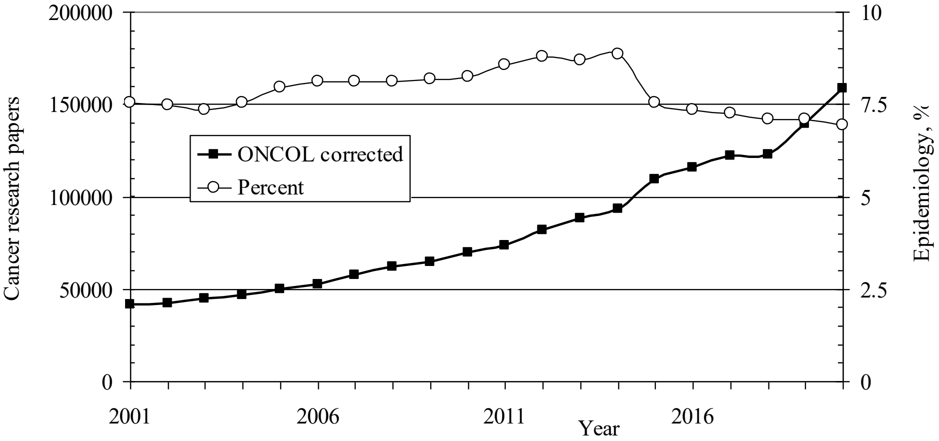


Figure 2. Outputs of cancer research papers (ONCOL, left axis) and percentage of epidemiology papers (right axis) in 2001-20.

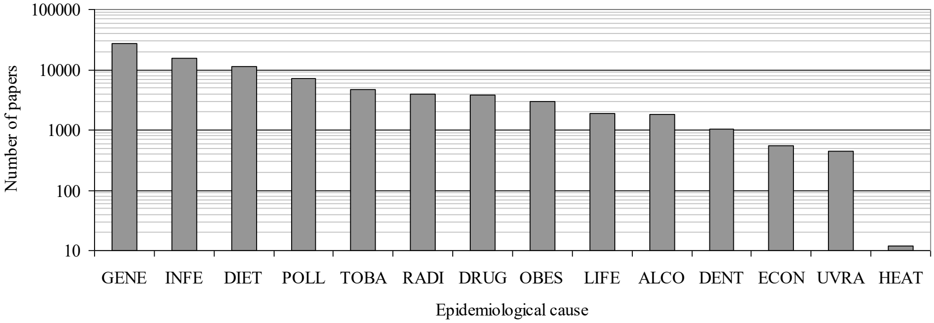


Figure 3. Outputs of cancer epidemiology papers on 14 causes, 2001-20; ***For tetragraphs, see Table 2***

The analysis of the 991 newspaper stories that we processed is shown in Figure 4, with solid columns for the seven Eastern European countries, and diagonal striped columns for the 13 Western European countries. Their output was dominated by three countries: the UK (285 stories), Belgium (118) and Ireland (111).

Comparison with Figure 3 shows that the newspapers concentrated much more on diet, lifestyle, pharmaceutical drugs, and tobacco than the researchers did. This is not surprising because these are subjects in which the readers can make choices. There are stories about genetics, which may be thought to generate the residual risks of cancer not specified in Table 1, but even in Western Europe they only account for just over 8% of the total, less than the 17% of research (Table 5). In Eastern Europe the disparity is greater: between 5.3% of newspaper stories and 28% of research outputs. Figure 5 shows a direct comparison of the results for both Eastern and Western Europe's news stories, and their research outputs.

Table 5. Percentages of all cancer epidemiology papers in eight continental regions that are in the 14 listed subject areas (see Table 2 for codes). Highest values in bold, second highest values are underscored.

Continent	*Africa*	*Asia*	*China*	*EEur*	*WEur*	*LatAmC*	*N Am*	*Oceania*	*World*
GENE	15.5	**17.9**	**34.4**	**28.1**	**16.6**	17.3	**16.5**	**20.4**	**18.7**
INFE	**26.3**	11.9	8.59	10.7	10.8	**21.7**	9.86	8.64	10.7
DIET	6.37	7.10	6.04	6.02	8.39	9.11	8.94	8.46	7.74
POLL	4.27	4.29	3.29	6.58	5.81	4.93	5.76	5.52	4.97
TOBA	2.55	3.61	2.41	3.47	3.00	2.96	4.17	3.09	3.26
RADI	1.86	2.32	1.50	4.24	3.35	1.26	2.88	3.00	2.74
DRUG	2.11	1.03	1.80	2.09	2.87	1.80	3.62	2.90	2.68
OBES	1.14	1.70	1.35	1.80	2.00	2.16	2.73	3.19	2.07
LIFE	1.14	0.72	0.54	1.29	1.60	1.26	1.71	2.53	1.29
ALCO	0.79	1.40	1.01	1.63	1.40	1.70	1.36	1.65	1.24
DENT	0.79	1.02	0.55	0.59	0.72	2.08	0.53	0.66	0.71
ECON	0.17	0.26	0.11	0.35	0.41	0.44	0.48	0.43	0.38
UVRA	0.07	0.11	0.06	0.18	0.38	0.22	0.41	1.34	0.30
HEAT	0.10	0.00	0.02	0.00	0.02	0.02	0.00	0.02	0.01

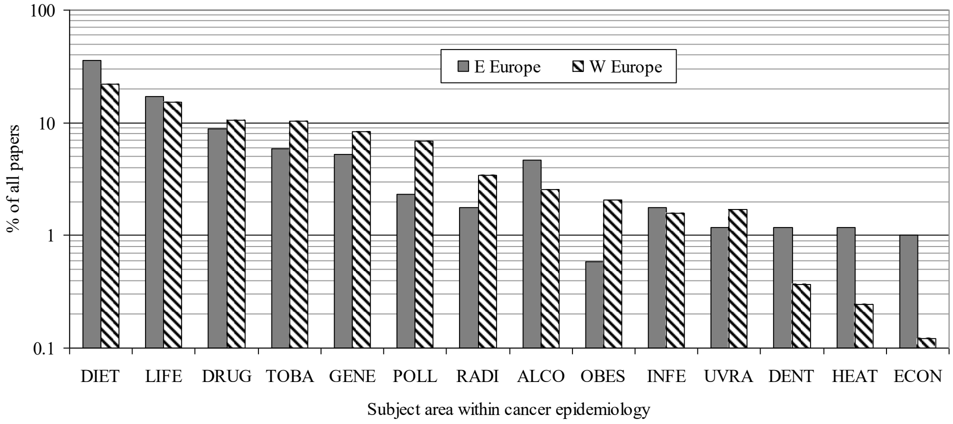


Figure 4. Chart of the percentage contributions to cancer epidemiology of 14 subject groups of European newspaper stories, 2002-13, integer counts. ***For tetragraphs, see Table 2.***

The IHME data on the causes of cancer in different continents are available, and some results for 2010 are shown in Table 6 (GBD Compare | IHME Viz Hub (healthdata.org)). These data show that there are some big differences between continents on which causes are the most important with regard to the development (or the reduction) of cancer.

Table 6. Percentage contributions to the causes of cancer for eight world regions in 2010, according to the IHME. ***Tetragraph codes are in Table 2. Regions are specified in Table 3.***

	Africa	*Asia*	*China*	*E Eur*	*W Eur*	*Lat Am*	*N Amer*	*Oceania*	*World*
ALL	30	38	46	48	47	36	49	41	42
TOBA	10	22	30	30	29	17	31	21	24
DIET	3.9	5.6	6.9	6.2	5.8	4.9	5.3	6.3	5.7
ALCO	3.3	4.6	5.2	7.2	7.3	3.8	4.9	7.5	5.1
LIFE	10	4.8	3.3	3.7	2.8	7.1	2.9	2.8	4.1
OBES	3.1	2.7	3.2	6.0	5.2	4.7	6.8	6.1	3.9
POLL	2.3	3.8	6.8	2.9	2.3	1.7	1.5	0.6	3.7

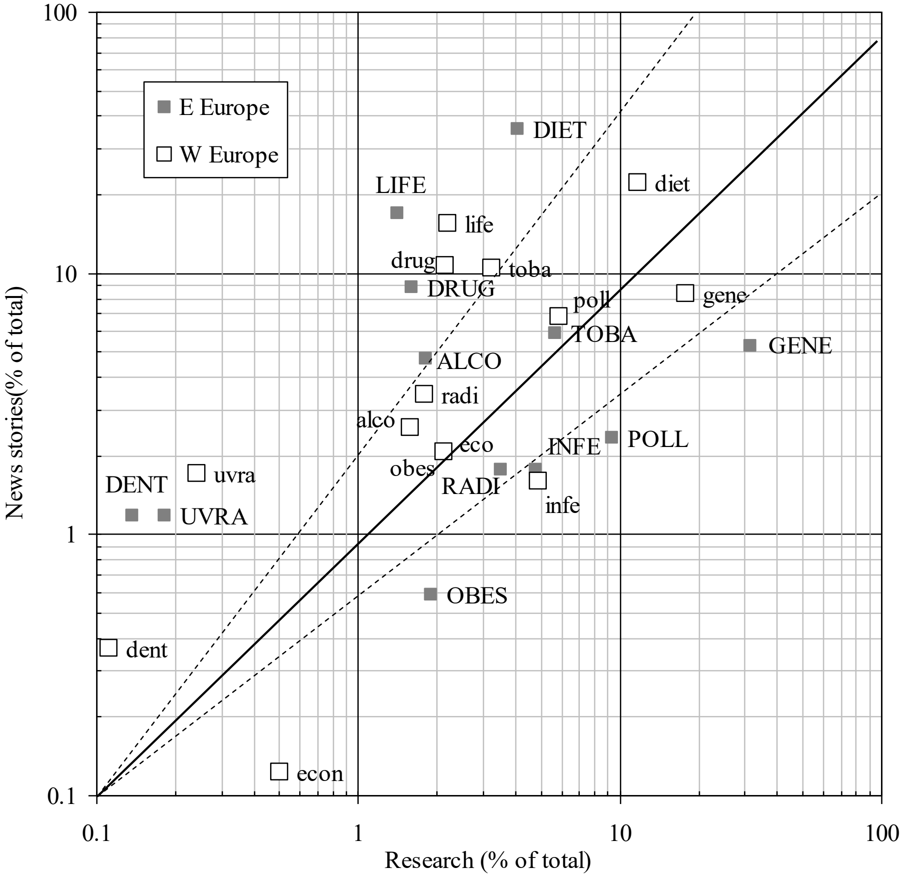


Figure 5. Comparison of newspaper coverage of the different subject areas of cancer epidemiology for Eastern Europe (solid squares, upper case labels) and for Western Europe (open squares, lower case labels), 2002-13, with their research outputs, 2001-20. Log-log scales. Solid diagonal line: equivalent percentages; dashed lines: news coverage five times or one fifth percentages of research. ***For tetragraph codes, see Table 2.***

The individual causes (from tobacco use to occupational causes = OCCU) add to more than the total, probably because of double counting with some causes combining to increase the risk of cancer. Tobacco as a cause of cancer varies greatly, from 31% in North America to only 10% in Africa. Alcohol has the biggest effects in Oceania and Europe, but much less in Africa, many of whose countries are Muslim-majority where alcohol is not freely available. However, Africa suffers more from lifestyle choices, principally unsafe sex. Obesity has the largest effect in North America, and the least in Asia (but not in China). China suffers mainly from its diet, and from pollution. So, these data tend to confirm some subjective views of the world.

It is possible to make a comparison of these causes in the eight continental regions with research data (taken from Table 4) for four of the subject areas listed in Table 1, tobacco, alcohol, diet and obesity. This is intended to show whether the researchers are focusing their attention on the areas deemed to be of importance to the control of cancer in the eight continents. The correlation is positive for all four; it is highest for obesity and lowest for diet, probably because dietary data are not so readily available as data for the other three.

Discussion

In this work, we report (what we think is for the first time) a comparison between the causes of cancer, the amount of research on each of the main risk factors, and the extent of their news coverage in Europe. News coverage is important because several of the causes of cancer are to some extent under readers' control, and therefore may be influenced by the stories that they read. It was clear from these data that the journalists had selected cancer research stories, not as an unbiased sample but rather ones that they thought would interest their readers. Thus, of the 74 stories about genetics, the largest number (27, or 36%) were about breast cancer, which is always a popular subject (Lewison *et al.*, 2008).

The 14 cancer risk factors are somewhat different from each other. Genetics (GENE) is the dominant subject area in most of them, but infection (INFE) is top in both Africa and Latin America because of the papilloma virus and its effect on cervical cancer, which is less well screened and treated there compared with Western Europe and North America (Vaccarella *et al.*, 2017). Pollution (POLL) is evidently a subject for a lot of research in Eastern Europe, as is smoking (TOBA). There is relatively little research on this in China; this may reflect the policy of the Chinese government, which derives very substantial tax revenues from cigarettes (Li & Lewison, 2020). The interest of the countries of Eastern Europe in radiation (RADI) is probably attributable to the long-term health effects from the Chernobyl nuclear accident in April 1986 (Ivanov *et al.*, 2012. Rivkind *et al.,* 2020). It is also notable that obesity (OBES) is of particular interest in Oceania (mainly Australia) and the rich countries of North America and Western Europe. Australia is also the most concerned with research on ultraviolet radiation (UVRA) because of its high incidence of solar-induced skin cancer (McLoone *et al.*, 2014; McMeniman *et al.*, 2020). We added the subject areas of dental issues and hot drinks because of the belief that dental health has wider effects on bodily health and the risk of various manifestations of cancer (Chen *et al.*, 2020; Kawasaki *et al.*, 2020; Wu *et al.*, 2021), and the problems, particularly in China, with oesophageal cancer which can result from the drinking of very hot tea (Yu *et al.*, 2018; Yu et al., Li & Lewison, 2020).

The impact of such modifiable risk factors and the disease burden of cancer, however, may not be entirely understood at the population level until there is a clear communication and awareness strategy from the government. Such a distinction between policy and action was made in a recent study that showed that government strategic plans focused on the treatment of diabetes rather than its prevention in Cyprus, Iceland, Luxembourg, Malta and Montenegro (Cuschieri *et al.* 2021). Furthermore, the importance of translating burden of disease studies for use in policy-making was recently raised with efforts concentrated in producing a knowledge translation framework (Pallari *et al*, 2020). To support these efforts the European

COST Action burden of disease network was set up. It is intended to integrate technical initiatives in the development and use of these metrics into findings relevant to stakeholders. It will also strengthen the capacity for cross-country collaboration, as reported by Devleesschauwer (2020). To this end, and to effect such a change in health communication and policy-making, data visualisation tools may enhance the engagement of decision-makers and relevant stakeholders, as a disconnect between research and public health has been reported (Lundkvist *et al.,* 2021).

Strengths and limitations

The study involved the analysis of voluminous research outputs at the global level, the comparative assessment using burden of disease data from IHME and examination of news-media reports. We were fortunate to have developed such a database of newspaper stories from so many different countries, albeit confined to Europe. Additionally, we have drawn comparisons between the countries concerned from the research outputs and newspapers with data on the causes of cancers as part of an effort in linking these data and identifying gaps in public health communication and informing policy-makers on how to reduce the burden caused by cancer in these countries.

The study has several limitations. Because of time, the file of 144,597 cancer epidemiology papers was a "virtual" one, and the analysis was based only on the WoS software. The details of the papers were not actually downloaded to file, which would have allowed us to examine the fractional counts of countries, which provide a fairer assessment of their individual contributions to the research. It would also have allowed us to link the various causes, *e.g.*, diet, alcohol use, to the different manifestations of cancer that are particularly affected, which might help researchers to focus their work better. We have done this only for two examples, but there must be many others besides the link between smoking and lung cancer, which is one of the best-established. The allocation of papers to causes that we performed has similarities with the topic modelling process in which documents are classified in multiple subject areas based on text mining using natural language processing techniques, although in this paper we manually examined samples of the papers and used our judgement to classify them through a complex iterative process.

Implications for further research

There is some evidence that cancer researchers in different continental regions are studying the subjects that are the most important for their influence on the risks. This is only an association, and it would be necessary to carry out a survey of some of the leading researchers to establish whether they are aware of the relative importance of the different causes, and whether this knowledge had influenced their choice of research topics. It would also be useful to ask funders of such research whether they try to support the research that was most likely to reduce the cancer burden in their countries, or whether their allocation of funds was based only on their perceptions of the potential quality of the work. Additionally, mapping out the factors based on a country analysis may provide useful results that are of value to local policy-makers.

Conclusions

Epidemiological research into the causes and prevention of cancer only represents a small percentage of the total, despite this being more cost-effective than treatment once the disease is established in a patient. However, it appears to be targeted at the most important causes in different continents. Although the prevention of cancer depends heavily on governmental public health campaigns (notably to control smoking), the lifestyle choices of members of the public also make a difference. These can be influenced by information about new research findings that are reported in the news-media. Within Europe, it seems that these are chosen to reflect the topics that are likely to have the most interest for their readers, which is helpful.

Acknowledgment

We are grateful to the many research assistants and volunteers who worked to analyse the European newspaper stories at King's College London. This project was led and managed by EP, now at University College London, and was supported by the European Commission's Seventh Framework Programme (grant no. EC/FP7/602536).

References

Begum M, Lewison G, Lawler M & Sullivan R. (2018) Mapping the European cancer research landscape: An evidence base for national and Pan-European research and funding. *European Journal of Cancer*, 100: 75-84.

Boffetta,P. (2007) Molecular cancer epidemiology: a tale of >3842 publications. *Carcinogenesis*, 28(8), 1621.

Chen Y, Zhu BL, Wu CC, Lin RF & Zhang X. (2020) Periodontal Disease and Tooth Loss Are Associated with Lung Cancer Risk. *BioMed Research International.* 5107696 DOI: 10.1155/2020/5107696

Cuschieri S, Pallari E, Terzic N, *et al.* (2021) Mapping the burden of diabetes in five small countries in Europe and setting the agenda for health policy and strategic action. *Health Research Policy and Systems.*19:43 DOI: 10.1186/s12961-020-00665-y

Devleesschauwer, B. (2020) European burden of disease network: strengthening the collaboration. *European Journal of Public Health,* 30(1): 2-3.

Flor LS, Reitsma MB, Gupta V et al. (2021) The effects of tobacco control policies on global smoking prevalence. *Nature Medicine*, 21 January. DOI: 10.1038/s41591-020-01210-8

Ginsberg GM, Edejer TTT, Lauer JA & Sepulveda C. (2009) Screening, prevention and treatment of cervical cancer - A global and regional generalized cost-effectiveness analysis. *Vaccine*, 27(43): 6060-6079

Hurry M, Eccleston A, Dyer M & Hoskins P. (2020) Canadian cost-effectiveness model of BRCA-driven surgical prevention of breast/ovarian cancers compared to treatment if cancer develops. *International Journal of Technology Assessment in Health Care*, 36(2) 104-112.

Ivanov VK, Kashcheev VV, Chekin SY *et al.* (2012) Radiation-epidemiological studies of thyroid cancer incidence in Russia after the Chernobyl accident (estimation of radiation risks, 1991-2008 follow-up period). *Radiation Protection Dosimetry*, 151(3): 489-499.

Kawasaki M, Ikeda Y, Ikeda E *et al.* (2020) Oral infectious bacteria in dental plaque and saliva as risk factors in patients with esophageal cancer. *Cancer*, DOI: 10.1002/cncr.33316

Lewison G. (1999) The definition and calibration of biomedical subfields. *Scientometrics*, 46(3): 529-537

Lewison G. Tootell S, Roe P & Sullivan R. (2008) How do the media report cancer research? A study of the UK's BBC website. *British Journal of Cancer*, 99: 569-576

Li A & Lewison G. (2020) Chinese Cancer Research in 2009-18 and the Disease Burden. *Cancer Management and Research*, 12: 5031-5040.

Lin MQ, Lian CL, Zhou P *et al.* (2020) Analysis of the Trends in Publications on Clinical Cancer Research in Mainland China from the Surveillance, Epidemiology, and End Results (SEER) Database: Bibliometric Study. *JMIR Medical Informatics,* 8(11): e21931.

Lundkvist, A., El-Khatib, Z., Kalra, N., *et al.* (2021). Policy-makers' views on translating burden of disease estimates in health policies: bridging the gap through data visualization. *Archives of Public Health*, *79*(1), pp.1-11.

McLoone JK, Meiser B, Karatas J *et al.* (2014) Perceptions of melanoma risk among Australian adolescents: barriers to sun protection and recommendations for improvement. *Australian and New Zealand Journal of Public Health*, 38(4): 321-325

McMenimanEK, Duffy DL, Jagirdar K *et al.* (2020) The interplay of sun damage and genetic risk in Australian multiple and single primary melanoma cases and controls. *British Journal of Dermatology*, 183(2); 357-366

Pallari E & Lewison G. (2019) How Biomedical Research Can Inform Both Clinicians and the General Public. *Springer Handbook of Science and Technology Indicators*,(W Glänzel et al. eds.), Springer Handbooks https//doi.org/10.1007/978-2-030-02522-3_22; 581-607.

Pallari, E., Thomsen, ST & Hilderink, HBM. (2020). Knowledge translation and health policy for burden of disease. *European Journal of Public Health*, *30*(Supplement_5), pp.ckaa165-1380.

Rivkind N, Stepanenko V, Belukha I *et al*. (2020) Female breast cancer risk in Bryansk Oblast, Russia, following prolonged low dose rate exposure to radiation from the Chernobyl power station accident. *International Journal of Epidemiology*, 49(2): 448-456

Ugolini D, Puntoni R, Perera *et al*. (2007) A bibliometric analysis of scientific production in cancer molecular epidemiology. *Carcinogenesis*, 28(8), 1774-1779

Vaccarella S, Laversanne M, Ferlay J & Bray F. (2017) Cervical cancer in Africa, Latin America and the Caribbean and Asia: Regional inequalities and changing trends. *International Journal of Cancer*, 141(10): 1997-2001

Wu HD, Zhang JJ & Zhou BJ. (2021) Toothbrushing frequency and gastric and upper aerodigestive tract cancers risk: A meta-analysis. *European Journal of Clinical Investigation*, e13478. DOI: 10.1111/eci.13478

Yu CQ, Tang HJ, Guo Y *et al*. (2018) Hot Tea Consumption and its Interactions With Alcohol and Tobacco Use on the Risk for Esophageal Cancer: A Population-Based Cohort Study. *Annals of Internal Medicine*, 168(7): 489

Zimmermann MR, Vodicka E, Babigumira JB *et al*. (2017) Cost-effectiveness of cervical cancer screening and preventative cryotherapy at an HIV treatment clinic in Kenya. *Cost-Effectiveness and Resource Allocation*, 15:13 DOI: 10.1186/s12962-017-0075-6

Are we there yet? Analyzing scientific research related to COVID-19 drug repurposing

Namu Park[1], Hyeyoung Ryu[2], Ying Ding[3], Qi Yu[4], Yi Bu[5], Qi Wang[4], Jeremy J. Yang[6] and Min Song[7,*]

[1]*namupark@yonsei.ac.kr*
Department of Digital Analytics, Yonsei University, Seoul, South Korea

[2]*hyryu115@uw.edu*
The Information School, University of Washington, Seattle, WA, United States

[3]*ying.ding@ischool.utexas.edu*
School of Information, University of Texas, Austin, TX, United States

[4]*yuqi@sxmu.edu.cn, 389172356@qq.com*
School of Management, Shanxi Medical University, Taiyuan, Shanxi, China

[5]*buyi@pku.edu.cn*
Department of Information Management, Peking University, Beijing, China

[6]*jejyang@indiana.edu*
Department of Internal Medicine, School of Medicine, University of New Mexico, Albuquerque, NM, United States

[7]*min.song@yonsei.ac.kr*
Department of Library and Information Science, Yonsei University, Seoul, South Korea
Corresponding Author: Min Song (ORCID: 0000-0003-3255-1600, min.song@yonsei.ac.kr)

Abstract

Drug repurposing may be a pivotal means of fulfilling urgent needs for treatment of the novel coronavirus disease 2019 (COVID-19), but current studies on drug repurposing for COVID-19 seem to show a lack of consensus in their drug candidate focus. Using bibliometric methods in a non-expert perspective, in a review of 34 published articles on the COVID-19 and drug-repurposing, we investigated obvious and less obvious points of consensus on drug candidates. To establish these two types of consensus, we first implemented document clustering. Within a set of five clustered papers, we established an obvious consensus, relying solely on the occurrence of entities by using term frequency and inverse document frequency and a comparison of mentioned drugs, finding that remdesivir and chloroquine were discussed with a certain degree of agreement. For the less obvious consensus, we created a drug entity co-occurrence network to establish low-high centrality combinations to probe the crucial drugs found in article clustering that are not plainly apparent through the mere counting of the occurrence of drug entities occurrences. Lopinavir emerged as having possibly potent effects in spite of underuse, while the mainstream of studies focus more on drugs such as chloroquine that enjoy explicit consent. Using an entitymetrics perspective, we expect that our research will support investigations of drug repurposing, expediting the process of establishing treatment for COVID-19.

Introduction

In the context of the current novel coronavirus disease 2019 (COVID-19) pandemic and the pressing need for adequate treatment of it, it can be instructive to revisit the history of treatment discoveries during past pandemics. Ebola virus disease (EVD) is a complex infectious disease with a much higher mortality rate than COVID-19, killing an overall average of 50% and up to 90% of those infected (World Health Organization 2020). The first outbreak of EVD was in Sudan and

Congo in 1976, in a village near the Ebola River, and dozens of intermittent outbreaks followed throughout sub-Saharan Africa over the last 40 years. Between 2017 and 2018, the severe outbreaks of the disease in Congo led World Health Organization to declare EVD a world health emergency (Feldmann and Geisbert 2011). The first EVD vaccine, Ervebo, was only approved in the United States in December 2019 (FDA 2019), more than 40 years after the first outbreak, and still no antiviral drug has been approved by the FDA to treat it. More than 140 drugs have been repurposed for EVD, and there are reports of in vitro potency and in vivo effectiveness in animal models and clinical trials with EVD patients (Bai and Hsu 2019). Remdesivir, originally developed for treating hepatitis C, has been tested on EVD, but it failed in a recent clinical trial (Mulangu et al. 2019). However, remdesivir has been proven to effectively shorten recovery times for COVID-19 patients (Beigel et al. 2020). Recently, the FDA authorized the emergency use of remdesivir to treat severe COVID-19 (Dolin and Hirsch 2020). Unlike in earlier pandemics, an unprecedented worldwide scientific workforce has been brought together from universities, research labs, pharmaceutical companies, and organizations, exhibiting a laser focus on finding novel treatment for COVID-19 through drugs or vaccines.

Drug repurposing seeks new uses of existing drugs and is known to effectively shorten treatment development times (Pushpakom et al. 2019). Due to the urgent need to find a cure, drug repurposing has become a mainstream goal for COVID-19 research. It is thus crucial for us to understand the current status of research in this area as relates to COVID-19 and in particular to establish whether scientists have reached any consensus on a list of candidate drugs with potential for COVID-19.

iWe retrieved 34 papers on COVID-19 literature from Kaggle, which is the world's largest data science community, with the criterion that each used scientific method to propose a list of drug-repurposing candidates for COVID-19. By analyzing the resulting list of recommended candidate drugs for COVID-19 using bibliometric methods, we demonstrated a lack of consensus among drug repurposing literatures. Specifically, we observed two types of consensus: obvious and indistinct consensus.

Our research constitutes an early and objective investigation of scientific consensus without direct input from active researchers, relying solely on methods of data science for ease and speed of research. In short, this study outlined the key scientific outputs of drug repurposing for COVID-19 through the use of entitymetrics (Ding et al. 2013) that studies how drug and disease entities affect knowledge transfer and impact different fields or subjects.

Related Work

Drug Repurposing

Drug repositioning, also called drug repurposing or drug reprofiling, is a common term in the drug discovery context where new therapeutic opportunities for existing drugs are sought (Doan et al. 2011). This method is attracting more attention at present due to the advent of COVID-19, but it has been in widespread use since the beginning of the twenty-first century. In 2006, a study of drug repositioning was conducted that used a computational approach (Li et al. 2006), and 3 years later, in silico compound profiling was used to broaden the knowledge map of drug repurposing (Dubus et al. 2009). After the onset of the global pandemic in 2020, the drug-repositioning literature grew rapidly. The main reason for this is that since SARS-CoV-2, which causes COVID-19, is a type of a coronavirus, researchers expected that drug repurposing was a plausible method for finding treatment (Altay et al. 2020).

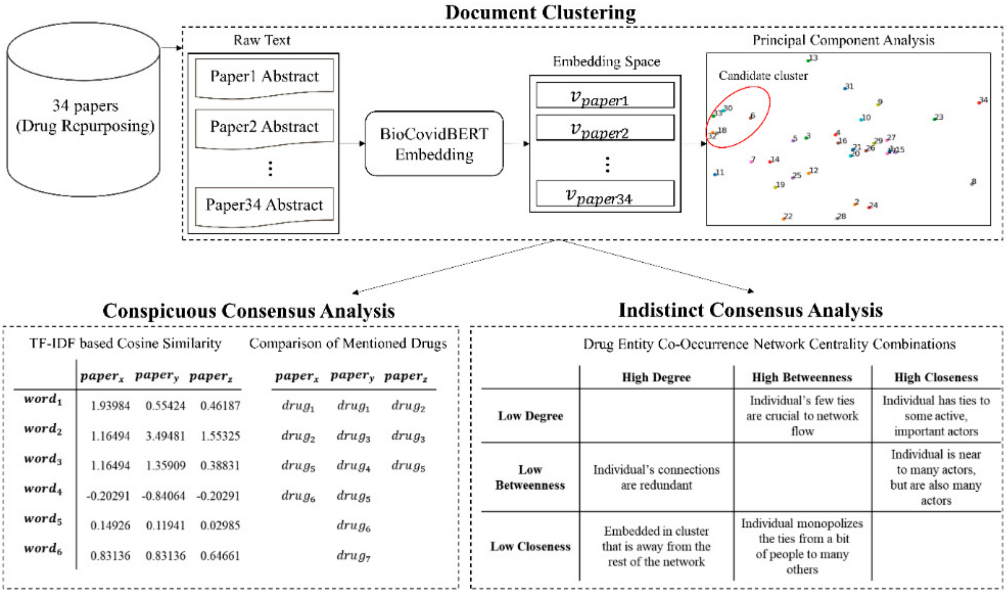
Consensus Check

There has been little treatment of the scientific consensus on COVID-19 treatment or any other subject because pursuing such research is time consuming and requires a certain level of domain knowledge. The most best-known research on scientific consensus, that of Cook et al. (2013), investigated the consensus on anthropogenic global warming . Its goal was to identify the evolution of this consensus, and it indicated that an increasing number of publications accepted the consensus. As a part of the citizen science project, volunteers with domain knowledge manually rated each publication's level of endorsement, adopting a criterion set by the authors. Nearly 25,000 abstracts were rated in this experiment. However, because all of the ratings were assigned manually, the work was time consuming, and the different volunteers had different backgrounds and perspectives, which likely affected their ratings. This lack could lead to doubt regarding the results and may damage the reliability of the conclusions. Therefore, we suggest a streamlined method of observing consensus without resorting to potentially biased opinions from experts.

Proposed Approach

We selected papers from PubMed that proposed a drug candidate for repurposing using the following search query:

drug candidate or repurpos* or reposition* or re-purpos* or re-position**



TF-IDF based Cosine Similarity

	$paper_x$	$paper_y$	$paper_z$
$word_1$	1.93984	0.55424	0.46187
$word_2$	1.16494	3.49481	1.55325
$word_3$	1.16494	1.35909	0.38831
$word_4$	-0.20291	-0.84064	-0.20291
$word_5$	0.14926	0.11941	0.02985
$word_6$	0.83136	0.83136	0.64661

Comparison of Mentioned Drugs

$paper_x$	$paper_y$	$paper_z$
$drug_1$	$drug_1$	$drug_2$
$drug_2$	$drug_3$	$drug_3$
$drug_5$	$drug_4$	$drug_5$
$drug_6$	$drug_5$	
	$drug_6$	
	$drug_7$	

Drug Entity Co-Occurrence Network Centrality Combinations

	High Degree	High Betweenness	High Closeness
Low Degree		Individual's few ties are crucial to network flow	Individual has ties to some active, important actors
Low Betweenness	Individual's connections are redundant		Individual is near to many actors, but are also many actors
Low Closeness	Embedded in cluster that is away from the rest of the network	Individual monopolizes the ties from a bit of people to many others	

Figure 1. Overview of the adopted approach

56 publications were found, among which we removed those that did not propose any repurposed drugs in their results parts. As a result, 34 papers were left. Our study had two main steps: document clustering and consensus analysis. Document clustering is the process of finding the set

of candidate publications for review. Using the candidate documents established in the first stage, we conducted two analytical stages to check the consensus exhibited among them. Conspicuous consensus analysis involves methods that rely solely on term frequency and inverse document frequency (TF-IDF) and a comparison of the mention of certain items. It checks obvious consensus among candidate documents. Indistinct consensus analysis, by contrast, targets implicit agreement among documents, using a method that assesses the occurrence of entities and different types of network centrality. Figure 1 illustrates the overall analytical process.

Document Clustering

First, we presumed that the abstract of an article is the summary of the entire document, and all essential information is always included in it. We extracted the abstracts of 34 papers and then plotted them in vector space.

The word-embedding model used in this experiment was BioCovidBERT developed by Tonneau (2020), a fine-tuned version of BioBERT, created by Lee et al. (2020). BioCovidBERT was trained on a preprocessed COVID-19 dataset accessible in the COVID-19 Open Research Dataset Challenge (Kaggle 2020). It was created with the help of several leading research groups in response to the COVID-19 pandemic, and over 181,000 coronavirus-related publications are included in the data of the COVID-19 Open Research Dataset Challenge. We assumed that BioCovidBERT, on this rich resource, could effectively represent each document as a vector while preserving their semantics, especially for words related to COVID-19. After document embedding, each paper was represented as a 1,024-dimensional vector. To check the similarity of the documents, we used Kernel Principal Component Analysis (KernelPCA), which can effectively reduce the dimensions of non-linear high-dimensional data. With KernelPCA, we visualized each document vector in a 2- and 3-dimensional space, and K-means clustering was used to elaborate groups of vectors. Among the derived clusters, we selected the one with the shortest intra-class distance and the longest inter-class distance, meaning that its components are self-similar, and the differences between points from other clusters are relatively high. The components of the selected cluster are determined as candidate documents and considered for further analysis in the next phase.

Conspicuous Consensus Analysis

With the candidate documents, two experiments were conducted for further investigation. To begin with, we used the TF-IDF for each candidate document. The following equation is the formula for TF-IDF, where N refers to the number of documents in the corpus:

$$TFIDF_{word,doc} = tf_{word,doc} * \log\left(\frac{N}{df_{word}}\right)$$

TF-IDF is a scalar value and is used to determine how important a word is in a document. TF gives the number of appearances of the target word in each document, and IDF represents whether it is used commonly or rarely for each document. For example, the word “the” appears commonly in most document, so df_{the} is relatively high, but inverse document frequency is low. Therefore, a high TF-IDF value indicates that the target word is rarely used in other documents but frequently appears in the target document.

For a more complete analysis, we extracted lists of all drugs mentioned in each paper and checked whether there were intersections between the drugs. This was done using PubTator Entity Tagger (U.S. National Library of Medicine and National Center for Biotechnology Information 2020) and with the help of the COVID-19 Drug and Gene Set Library developed by the Mount Sinai Health System (Icahn School of Medicine at Mount Sinai 2020), and we double-checked whether the extracted entities are included in the drug set. A total of 147 drugs were mentioned in 34 papers.

Indistinct Consensus Analysis

We formed drug entity co-occurrence networks for the clustered papers by setting the nodes as entity instances and the edges as the number of co-occurrences between the entity instances, making an entity co-occurrence network. In this network, we calculated the degree, betweenness, and closeness centrality values for each node (i.e., entity instance). After the logarithmic values for degree and closeness centrality values and the square roots of the betweenness centrality values were calculated, we cut the values into terciles, labeled as low, medium, and high. Using the values in the low tercile for one centrality and those in the high tercile for another, we created a low–high centrality combination for each entity. This combination indicates hidden gems within the entity network that cannot be observed from a direct observation of the centrality combinations.

In Table 1, we list the features of each centrality combination for the low–high centrality combinations (Zhang and Luo 2017). The instances in each entity co-occurrence network in the low–high centrality combination table could affect the network in a non-obvious but pivotal way. However, it should be noted that not all instances that occur in the low–high centrality combination table have low-profile importance, and that those that do appear in one of the following four low–high centrality combination groups: (a) low degree–high betweenness, (b) low degree–high closeness, (c) low closeness–high degree, and (d) low closeness–high betweenness.

Table 1. Low–high centrality combination features

	High Degree	*High Betweenness*	*High Closeness*
Low Degree	-	An individual's few ties are crucial for network flow.	An individual has ties to some important actors
Low Betweenness	An individual's connections are redundant and communication bypasses him/her	-	The network may hold many paths where an individual is near many actors, but so are many others.
Low Closeness	An individual is embedded in a cluster that is away from the rest of the network.	An individual can monopolize the ties of a few people to many others.	-

Results

Document Clustering

Figure 2 on the next page represents document embedding in a 2-dimensional space. At first glance, it is not easy to observe the existence of meaningful clusters or find a consensus among the 34 drug repurposing papers. To develop our investigation, we used the K-means clustering algorithm, an unsupervised machine learning approach that is widely used for cluster detection. Our goal was to check potential clusters within this 2-dimensional space, where horizontal and vertical axes represent the principal components. In other words, we assumed that if there are to be similar points among some of the papers, they would form a cluster in vector space, in which the components may share elements. Before applying the K-means method, it was necessary to develop a process to derive the optimal number of clusters. Using the elbow method and silhouette scoring, we discovered that three is an optimal number for our clusters. The results of our clustering are given in Figure 3-a, where the six dots on the left side of the 2-D plane form a relatively meaningful cluster, as the inter-class distance is longer there than in the other clusters, and intra-class distance is shorter. Therefore, we anticipated that documents corresponding to the points on the left side of the 2-D plane (6, 7, 11, 18, 30, 32, and 33) have the highest probability of addressing issues common to all 34 papers. We also checked KernelPCA in 3-dimensional space to improve the validity of the candidate documents (Figure 3-b). We found that documents 7 and 11 had low similarity in 3-dimensional space, although they seemed were located close together in 2-dimensional space. Using these means, five documents (6, 18, 30, 32, and 33) were shortlisted for possible consensus publications.

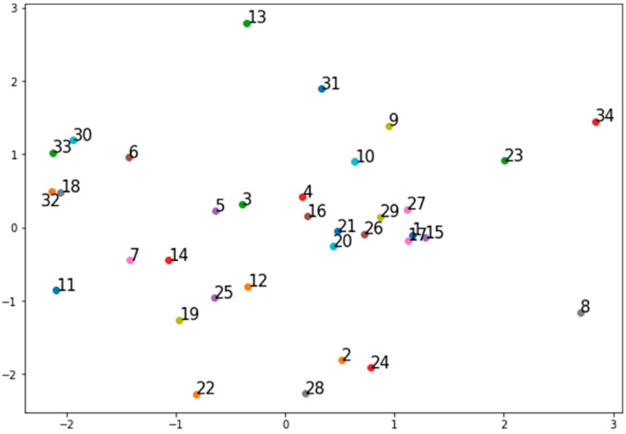
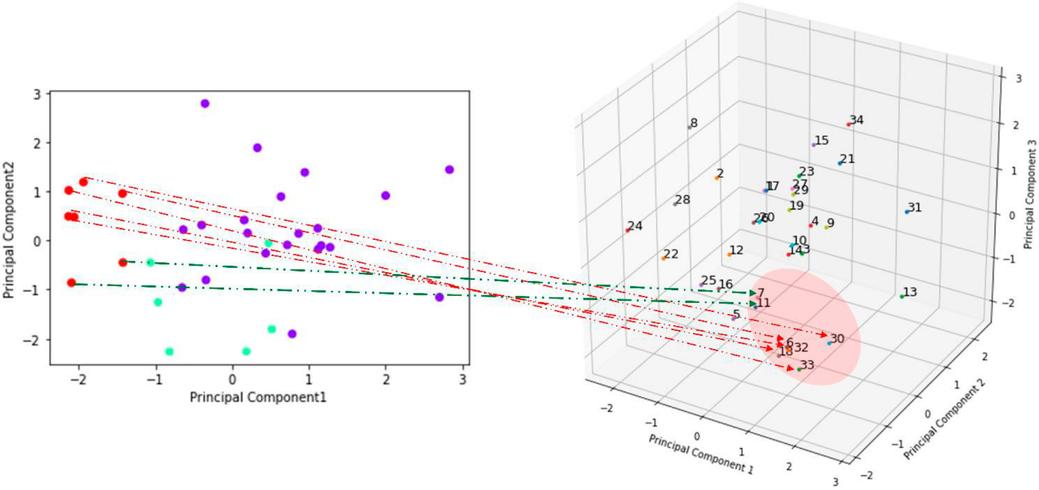


Figure 2. BioCovidBERT embedding of 34 papers projected in a 2-D space using Kernel PCA



(a) (b)

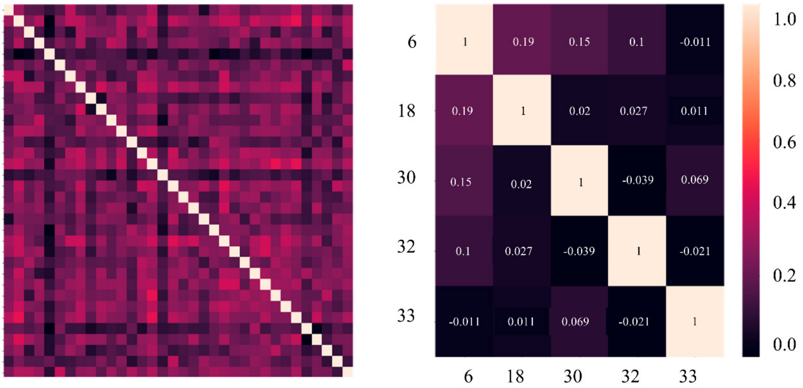
Figure 3. Projection of document embedding in 2-D, 3-D space

Conspicuous Consensus Analysis

Table 2. Part of the TF-IDF matrix for the candidate documents

	Paper 6	**Paper 18**	**Paper 30**	**Paper 32**	**Paper 33**
Receptor	0.00000	2.78645	0.00000	0.34831	0.34831
Hydroxychloroquine	0.00000	7.34265	1.22378	0.00000	40.38459
Plasma	13.46153	0.00000	0.00000	0.00000	0.00000
ACE2	0.00000	3.27324	0.00000	0.00000	2.45493
Chloroquine	0.00000	45.38988	2.32769	1.16384	29.67800
Remdesivir	0.00000	0.43532	0.87064	3.48254	0.87064
Protease	0.00000	0.30748	0.00000	5.22724	0.61497
Pneumonia	0.00000	0.00000	1.17260	0.39087	0.00000
Azithromycin	0.00000	0.00000	0.00000	0.00000	21.08615
Lopinavir	0.53063	0.00000	1.06126	1.59188	3.18377

After removing stop words, we identified the 1,374 most frequently used words from 34 papers to generate a TF-IDF matrix. Table 2 presents part of this matrix for the five candidate documents, and certain several characteristics of each paper can be observed. In Paper 6 (Da Silva 2020), entitled "Convalescent plasma: A possible treatment of COVID-19 in India," convalescent plasma, which can be collected from an infected individual, is proposed as a treatment to COVID-19. The TF-IDF for plasma in Paper 6 is high, but it is not mentioned in other four documents. In Paper 18 (Devaux et al. 2020), it is reported that chloroquine interferes with ACE2, a protein on the surface of the cell that is known to be a pathway for coronavirus penetration. The TF-IDF values for chloroquine and ACE2 are relatively higher than its TF-IDF values from other documents. Thus, our TF-IDF matrix reflects each document's main ideas quite well. Following this assumption, we calculated the cosine similarity for each TF-IDF vector from each publication (Fig. 4).



(a) Cosine similarity of TF-IDF vectors from all 34 papers

(b) Cosine similarity of TF-IDF vectors from 5 candidate papers

Figure 4. Heatmap of cosine similarity based on the TF-IDF vectors

The average cosine similarity of TF-IDF vectors was 0.1556 when calculated with all 34 papers, whereas the similarity values for the five candidate papers were lower, with an average of 0.1133. The fact that all five candidate documents, which formed the most meaningful cluster in the clustering section, showed a lower level for TF-IDF vector similarity implies two things.

A first is that there is information loss when the dimensions of document embeddings are reduced. On this view, some common issues may appear to be shared in a 2- or 3-dimensional space, but there is no consensus in reality. A second is that, if there was no information loss, the cross sections among these publications may be too trivial. To be specific, the words that contribute to the clustering may not be about the 1,374 core words we used to generate TF-IDF vectors. If some critical information were shared, shared vocabulary used to present that discovery would result in a high similarity in TF-IDF vectors. Overall, only eight drugs were mentioned more than twice: lopinavir, hydroxychloroquine, Arbidol, chloroquine, ritonavir, remdesivir, favipiravir, and ribavirin. Remdesivir and chloroquine were each mentioned in four of the five candidate documents, followed by lopinavir, ribavirin, and ritonavir, which appeared three times each. Thus, remdesivir and chloroquine appear to enjoy a certain consensus, but there is still no broad agreement, as these documents discuss too wide of a range of drugs. Figure 5, presenting the network visualization of Table 3, which reports the list of drugs mentioned in the five publications, buttresses our point. The self-loops in Figure 5 highlight drugs that appear only in a single document. For example, potassium, amiodarone, and azithromycin are only mentioned in 33. The multiple self-loops observed in Figure 5 and the multitude of edges labeled remdesivir or chloroquine reinforces our claim that no strong consensus exists, except for these two drugs.

Table 3. Mentioned drugs list in five candidate papers

Paper 6	*Paper 18*	*Paper 30*	*Paper 32*	*Paper 33*
Arbidol	Chloroquine	Chloroquine	Adenosine	Amiodarone
Darunavir	Chlorpromazine	Lopinavir	Arbidol	Azithromycin
Favipiravir	Doxycycline	Remdesivir	Chloroquine	Chloroquine
Lopinavir	Hydroxychloroquine	Ribavirin	Disulfiram	Hydroxychloroquine
Methylprednisolone	Quinine	Ritonavir	Favipiravir	Potassium
Ritonavir	Remdesivir	Teicoplanin	Galidesivir	Remdesivir
-	Ribavirin	-	Lopinavir	-
-	Tyrosine	-	Nitazoxanide	-
-	-	-	Oseltamivir	-
-	-	-	Remdesivir	-
-	-	-	Ribavirin	-
-	-	-	Ritonavir	-
-	-	-	Tenofovir	-

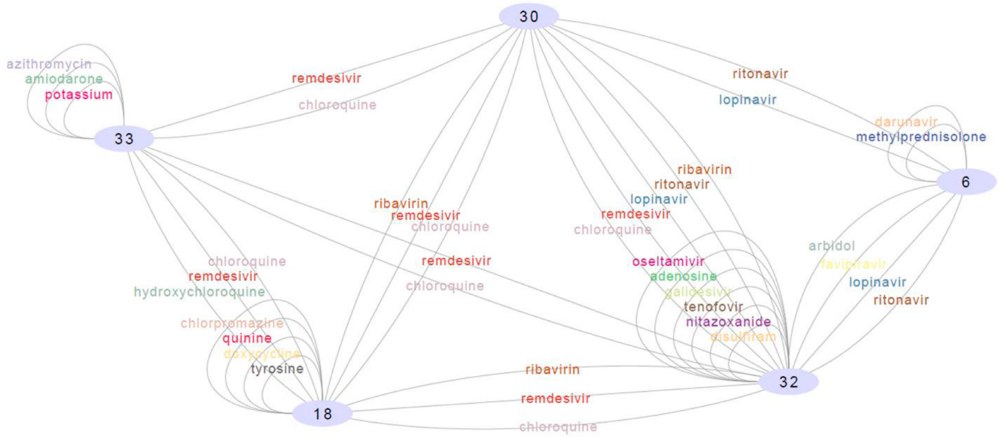


Figure 5. Multigraph network illustration of the mentioned drugs in the five candidate papers

Indistinct Consensus Analysis

In the drug entity co-occurrence network for the five clustered papers, chloroquine and teicoplanin were found to have high degree centrality scores, but they also had low betweenness and closeness centrality scores, which meant that their connections were redundant: important communication simply bypassed them, and they were embedded in a cluster that was distant from the rest of the network (Table 4). Lopinavir had few ties that were pivotal to network flow but had ties to other important drug instances in the network. This implies that lopinavir may be a hidden gem for the cluster formation of drug repositioning research in a high entropy situation.

Table 4. Drug entity co-occurrence network in low–high centrality combination for the five clustered papers

Drug	*High Degree*	*High Betweenness Centrality*	*High Closeness Centrality*
Low Degree	-	Lopinavir	Lopinavir
Low Betweenness	Chloroquine, Teicoplanin	-	-
Low Closeness	Chloroquine, Teicoplanin	-	-

Discussion

The data used to find conspicuous and indistinct consensus in treatments for COVID-19 were abstracts of COVID-19 drug repositioning papers published before April 15, 2020. To determine whether our predicted focusable drug was actually a target of COVID-19 drug repositioning research, we examined research trends before and after April 15, 2020, for the three drug instances found in low–high centrality combinations in the drug entity co-occurrence network. We investigated the number of research papers on each drug before and after April 15, 2020, on PubMed and calculated the rate of increase in the research by dividing the number of papers published after April 15, by the number of papers published before that date.

In Table 5, it is shown that all three drugs increased in the number of studies targeting them, but the higher research increase rate for chloroquine and teicoplanin should be noted. Although these two were not in the main part of the co-occurrence network, and important communication bypassed them, the number of studies conducted on these two were higher than those on lopinavir, which has few ties but ones that are crucial the network flow. The difference in research increase rate between lopinavir and chloroquine may not be significantly different, but a comparison of the absolute number of studies conducted shows that chloroquine was researched approximately 4.517 times more often than lopinavir, which indicates a notably low research focus on the latter. Thus, since the inconspicuous consensus analysis did not receive sufficient attention, probing this drug may help develop innovatory drug repositioning results.

Table 5. Lopinavir, chloroquine, and teicoplanin research trends before and after April 15, 2020

	Lopinavir	*Chloroquine*	*Teicoplanin*
Before April 15, 2020	11	47	1
After April 15, 2020	60	271	7
Research Increase Rate (# of Papers After April 15, 2020/# of Papers Before April 15, 2020)	5.455	5.766	7.0

Conclusion

In this research, we examined the consensus among publications on COVID-19 drug repurposing using two data science approaches. First, we derived several conspicuous features using document clustering, a TF-IDF matrix, and co-occurrence of drug entities. Among five papers that formed a cluster, we found that the cosine similarity for each publication's TF-IDF was low, and that only eight drugs were mentioned more than twice in this group. Remdesivir and chloroquine appeared in four out of the five clustered papers, which implied the conclusion that these drugs are the subjects of a certain agreement. Second, using co-occurrence network analysis, we conducted an additional experiment to determine an indistinct consensus. Lopinavir was the only drug to show high betweenness centrality, high closeness centrality, and low degree, which implies that it may have potential for repurposing. However, the rate of increase in research on chloroquine is still higher than that of lopinavir, which indicates that most publications are still concentrating on chloroquine.

In future work, we hope to extend our search for drug repurposing publications on the rapidly growing area of COVID-19. Applying these methods, we can continue to monitor what overlaps will emerge from the proposed repurposed drug candidates for COVID-19 from novel studies. These overlaps may eventually lead to a research consensus on an established list of repurposed drug candidates for COVID-19.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1A2C2002577). Qi Yu gratefully acknowledges financial supports from National Natural Science Foundation of China (Grant Number: 71573162) and the Shanxi Scholarship Council of China (Grant Number: HGKY2019057). We thank the support from NSF RAPID (2028717).

References

Altay, O., Mohammadi, E., Lam, S., Turkez, H., Boren, J., Nielsen, J., et al. (2020). Current status of COVID-19 therapies and drug repositioning applications. Iscience, 101303.

Bai, J. P. F., & Hsu, C. W. (2019). Drug repurposing for Ebola virus disease: Principles of consideration and the animal rule. *Journal of Pharmaceutical Sciences*, *108*(2), 798-806.

Beigel, J., Tomashek, K. M., Dodd, L. E., Mehta, A. K., Zingman, B. S., Kalil, A. C., et al. (2020). Remdesivir for the treatment of COVID-19: Preliminary report. *New England Journal of Medicine*, https://www.doi.org/10.1056/NEJMoa2007764

Cook, J., Nuccitelli, D., Green, S. A., Richardson, M., Winkler, B., Painting, R., et al. (2013). Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental Research Letters*, *8*(2), 024024.

Da Silva, J. A. T. (2020). Convalescent plasma: A possible treatment of COVID-19 in India. Medical journal, Armed Forces India, 76(2), 236–237. Advance online publication. https://doi.org/10.1016/j.mjafi.2020.04.006

Devaux, C. A., Rolain, J. M., Colson, P., & Raoult, D. (2020). New insights on the antiviral effects of chloroquine against coronavirus: what to expect for COVID-19?. International journal of antimicrobial agents, 105938.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., et al. (2013). Entitymetrics: Measuring the impact of entities. *PLoS ONE*, *8*(8), 1–14.

Doan, T. L., Pollastri, M., Walters, M. A., & Georg, G. I. (2011). The future of drug repositioning: Old drugs, new opportunities. In *Annual reports in medicinal chemistry* (Vol. 46, pp. 385-401). Academic Press. https://doi.org/10.1016/B978-0-12-386009-5.00004-7

Dolin, R., & Hirsch, M.S. (2020). Remdesivir - An important first step. New England Journal of Medicine, https://www.doi.org/10.1056/NEJMe2018715

Dubus, E., Ijjaali, I., Barberan, O., & Petitet, F. (2009). Drug repositioning using in silico compound profiling. *Future Medicinal Chemistry*, *1*(9), 1723-1736.

FDA. (2019). First FDA-approved vaccine for the prevention of Ebola virus disease, marking a critical milestone in public health preparedness and response. https://www.fda.gov/news-events/press-announcements/first-fda-approved-vaccine-prevention-ebola-virus-disease-marking-critical-milestone-public-health. Accessed 2 June 2020.

Feldmann, H., & Geisbert, T. W. (2011). Ebola haemorrhagic fever. *Lancet*, *377*(9768), 849-862.

Icahn School of Medicine at Mount Sinai. (2020). The COVID-19 Drug and Gene Set Library. https://amp.pharm.mssm.edu/covid19. (Accessed 3 June 2020.

Kaggle. (2020). COVID-19 Open Research Dataset Challenge (CORD-19) [Data file]. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. Accessed 2 June 2020.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Li, Y. Y., An, J., & Jones, S. J. (2006). A large-scale computational approach to drug repositioning. *Genome Informatics*, *17*(2), 239-247.

Mulangu, S., Dodd, L.E., Davey, Jr. R. T., Mbaya, O. T., Proschan, M., Mukadi, D., Manzo, M. L., et al. (2019). A randomized, controlled trial of Ebola virus disease therapeutics. *New England Journal of Medicine*, *381*, 2293-2303.

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2019). Drug repurposing: Progress, challenges and recommendations. *Nature Reviews Drug Discovery*, *18*(1), 41-58.

Tonneau, M. (2020). Covid-BERTs. https://github.com/manueltonneau/covid-berts. Accessed 2 June 2020.

U.S. National Library of Medicine, National Center for Biotechnology Information. (2020). Pubtator Central. https://www.ncbi.nlm.nih.gov/research/pubtator/. Accessed 3 June 2020.

World Health Organization (2020). Ebola virus disease: Key facts. https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease. Accessed 2 June 2020.

Zhang, J., & Luo, Y. (2017). Degree centrality, betweenness centrality, and closeness centrality in social network. In *2017 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017)*. Atlantis Press.

Scientific Awards: Women Laureates in STEM

Angayar Kumari Pavanasam[1] and Aparna Basu[2]

[1] *angikp@yahoo.com*
International College of Engineering and Management (UCLan UK) (Formerly), Muscat (Oman)

[1] aparnabasu.dr@gmail.com
Independent Researcher, Formerly at CSIR-NISTADS and South Asian University (India)

Abstract

Science, technology, engineering and math (STEM) fields are crucial for innovation capacity and life long learning in a rapidly changing world. But the gender gaps are so pronounced globally in STEM education and thereby in the workforce as well. Does this also reflect on the awards and recognitions bestowed on female scientists. This work attempts to examine the status of women laureates in terms of selected prestigious awards including Nobel prize, Fields medal, Abel prize and Turing award through scientometric analysis.

Introduction

Women are underrepresented in the fields of Science, Technology, Engineering and Mathematics (STEM), globally. In the last 50 years or so, the trend is improving but nevertheless the increase is not so sharp as to compensate for the challenges women face. In almost all parts of the world, a crude representation of the ratio of male to female population stands equal. The gap begins as early as in high school enrolments, though not sizable, but then widens in university degrees enrolment (particularly in STEM), and then on to pursuing postgraduate & research doctoral degrees. This, in turn, creates an imbalance in the national economy with regard to expenditures on women's education, and also leads to less representation of women in the general science-oriented workforce. Needless to say, there are far fewer women tackling global issues like climate change, renewable energy and so forth. In spite of barriers and the glass ceiling, some women do make it to the top in science fields. This necessitates to study the women's presence in the realm of prestigious academic awards.

There are a number of ways to assess the role of women in STEM and thereby examine the impact of their representation in the national economy (Beede et al. (2011)). This work tries to explore scientometric parameters of elite women achievers in terms of indicators like recipients of selected prestigious & coveted awards which include Nobel Prize (in Physics, Chemistry, Medicine or Physiology), Fields Medal & Abel Prize (both for Mathematics) and the Turing Award (Computer Science).

In the last 100+ years, only 28 women have received Nobel (119 years old), Fields Medal (84 years old), Abel Prize (17 years old) and Turing Award (53 years old) totally. Though it is known that women's representation is in these scientific awards, it is appalling to note that it is so low. It amounts to 3.5% of recipients' in Mathematics, 2% in Physics, 5% in Chemistry and 9% in Medicine or Physiology. If one looks into the geographical distribution, these women achievers in Mathematics are from Iran and US. Four female Physicists are from Canada, Poland each and two are from US whereas seven women from US, UK, France, Poland and Israel have received awards in the field of Chemistry. Twelve Laureates in the field of Physiology or Medicine field are from different countries viz., Australia, China, France, Germany, Italy, Norway and US amongst which 50% are from US. These statistics clearly indicate that the majority of the achievers are from US & EU either as a country of origin or adopted as emigrant citizens (nobelprize.org, abelprize.no, mathunion.org, amtur-

ing.acm.org).
There seems to be various factors that contribute to limited number of women achievers in STEM fields which include fewer women pursuing doctoral degrees in STEM, continuance of women doctoral degree holders as academicians / researchers, fewer publishing research works, receiving grants or establishing laboratories for a sustainable research environment, etc. All this together with breaks at these stages for family and elder care restrict a woman's opportunities to become achievers and this ultimately becomes a vicious circle.

As far as the authors know, particular characteristics of women in Nobel awardee lists have not been critically analyzed. Hence, this work focuses on a Scientometric analysis (through Web of Science / Scopus) of the scientific output of these women Nobel Laureates: Number of articles published, impact factors, and number of citations. The outcome of this analysis thus may help in revisiting the pathways & or policies related to women's empowerment in STEM with particular emphasis on cross-disciplinary activities, collaboration networks, and support toward exceptional creativity in scientific research.

Methodology

The research design was to collect the data of all elite awardees which included Nobel Laureates (Chemistry, Medicine & Physics), Abel Prize winners, Field Medalists, and Turing Awardees. Data regarding name, gender, year of award, date of birth and relevant information are taken from their respective websites (nobelprize.org, abelprize.no, mathunion.org, amturing.acm.org) and compiled for this study. For scientometric study, the data are drawn from the Scopus covering the years 2000-2018 in order to have reasonable sample distribution of both gender. To check the prevalence of gender disparity with regard to age and the scientific accomplishments, a non-parametric statistical analysis using t-test was done.

Gender asymmetry in prestigious awards

The growing importance of the issue of gender imbalance in STEM has led to discussion on number of defining factors operating over a period of time like leaky pipeline, crystal glass ceiling, scissors effect, impossible pursuit and overtaking model (Naldi et al (2004)). In STEM fields, making a critical breakthrough is always intriguing as it is an interplay of numerous factors including the foundational knowledge, consistent training, access to resources, cultural background, science of work (nature, theory, experimental), age, educational graph to list a few. In spite of major changes in higher education and in scientific career pursuits, gender imbalances still persists today especially in STEM fields.

Lack of use of women's skills and knowledge weigh heavily in STEM fields as shown in Fig. 1. The data has been presented here into two blocks (from the inception of the awards to 1970 as pre-1970 & 1971-2020 as post-1970) just for brevity and ease of interpretation. If we look at the data of gender distribution of the coveted awards, it clearly indicates an asymmetrical pattern across all fields as shown in Table 1.

Age

From around the time of Einstein, deliberations about age and productivity of scientists have taken place. It was during the early 20th century, when the majority of awardees were young , the average age was less than 50 across all fields. With time it has been constantly growing and interestingly in 2019, Physicist John B Goodenough received his Nobel Prize at the age of 97 contrary to the popular belief & saying by Einstein & Dirac that by age 30, a Physicist was effectively dead. Thus scientists receiving at the age above 50 has been on increase. The age at which scientists do their most significant work is also an important parameter when considering age and productivity.

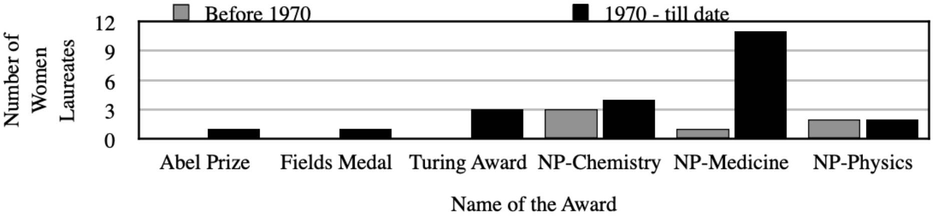


Figure 1. Women Laureates: Pre & Post 1970 status.

Table 1. Gender distribution across all fields

Award	*Field*	*Period*	*Total*	Ratio F/M
Abel Prize	Mathematics	2003-2020	22	5:95
Fields Medal	Mathematics	1936-2019	60	2:98
Turing Award	Computer Science	1966-2019	72	4:96
Nobel Prize	Chemistry	1901-2020	186	4:96
Nobel Prize	Medicine/Physiology	1901-2020	222	6:94
Nobel Prize	Physics	1901-2020	216	2:98

With the available data so far, the average age of women winning these awards are in 50-59 yrs band which appears to be similar to that of male awardees. Interestingly the age band has been increasing of late which is a welcome sign and might be attributed to return after break in career, increasing publications, availability & success of research grants, to name few factors.

It is also interesting to note that the age - productivity relationship is found to be field dependent and also the age at which award winning work was carried out (Table 2). It is only logical to infer that with time the odds (in terms of resources, time available, experiments) for researchers decrease and fall off precipitously after age 50 particularly in the fields of chemistry and physics which require more experimentation (Fones & Weinberg, 2011). In order to know, if there exists any difference in the ages at which these awards were received by men and women, the statistical t-test is done (Table 2). As seen in the table, the computed t-value is compared with the critical t-value at a confidence level of 95%. It is found that the computed value is less than the critical value, the result suggesting that there is no significant difference between the age of men and women at the time of winning the award.

Demographical Distribution

Out of 28 award recipients, if one looks at their locations at the time of receiving their awards, 14 of them are attached to USA based universities, whereas 5 from France (this includes Marie Curie receiving it in Physics & Chemistry), 2 from Israel, rest 1 from UK, Italy, Germany, Norway, China, Canada & Iran each. So, USA has more women laureates but which is quite similar to that of male recipients as well.

Field of Study

Even in 2020, the gender asymmetry is very much visible overall and across almost all fields. When we examine the data as pre & post 1970, there is no major sign of growth in terms of women awardees (Fig. 1). Before 1970, women never received Fields Medal and Turing award whereas the distribution of male:female of Nobel Prizes in Chemistry, Medicine and Physics, till date since inception, stands at 96:4, 95:5 and 98:2. It is relevant to note that Marie Curie is the only woman to receive two Nobel Prizes that too in different fields, so far. In the last 50 yrs (1970-2020), there is only a very marginal increase in the above fields (Chemistry, Medicine and Physics) and stands at 95:5, 91:9 & 98:2 as of now. The first woman to break the glass ceiling in Mathematics happened only recently i.e. at 2014 & 2019 viz., Maryam Mirzakhani being the first woman to receive Fields Medal in 2014 and Karen Uhlenbeck is the first woman to receive Abel Prize in 2019. So, the gender disparity is very much prevalent across all STEM fields. If one compares the data of gender gap in STEM vs NON-STEM (including Nobel Prize in Literature, Peace and Economic Sciences) fields, in STEM there is roughly only 1 woman for every 30 men whereas in NON-STEM the ratio seems to be encouraging at 1:10.

Award	*Field*	*Age Range*	*Female (avg)*	*Male (avg)*	t_{calc}	$t_{critical}$
Abel Prize	Mathematics	63-90	77	77	0.71	2.069
Fields Medal	Mathematics	31-40	37	36	0.92	2.160
Turing Award	Computer Science	50-79	65	57	0.99	2.035
Nobel Prize	Chemistry	43-98	51	56	0.35	2.007
Nobel Prize	Medicine/Physiology	46-85	63	58	0.28	2.007
Nobel Prize	Physics	36-96	54	59	0.02	2.004

Table 2. Comparison of average age of winners and statistical analysis data

Fractional Distribution

While much has been made of the Nobel's large gender imbalance, an analysis based on fractional counting reveals an even wider gap. An assessment of gender imbalance based on full counting, which doesn't account for prize share, reveals that women have received 20 (3.24%) of the 624 Nobel medals awarded since 1901. But an analysis based on fractional counting, considering prize share, finds that only 2.74% of the 334 sciences prizes have gone to women (natureindex.com). Out of 28 women who have received prizes across Chemistry, Medicine and Physics fields, 21% of them received full share, 32% of them received half, 18% received one-third share and 29% received quarter share of the prize value.

Scientometric Analysis

The data for scientometric analysis are drawn from the Scopus database for the period 2000 - 2018. The collected data includes number of citations of the individual authors, number of documents authored by the awardees and h-index of these awardees. While looking at these data, it is not surprising to see the asymmetric pattern given the low number of women awardees (Fig. 2) (Pavanasam & Basu, 2019).

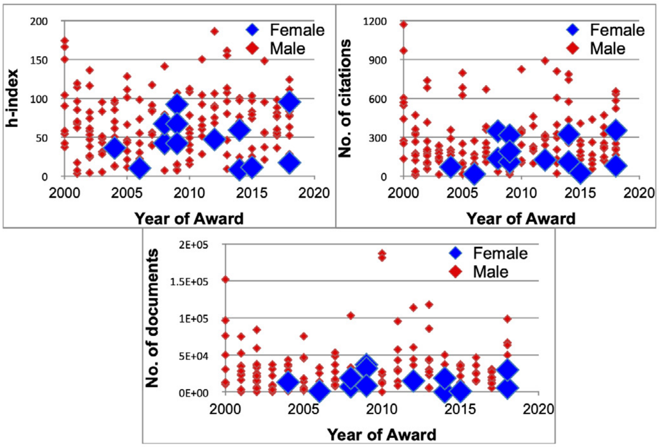
For a better understanding, *t*-test was carried out for all 3 parameters viz., total citations of the awardees, total documents created by the authors and h-index (Table 4). It is found that for all

3 data, the calculated *t*-value is lesser than the corresponding critical *t*-value suggesting there is not enough evidence for a significant difference between scientific achievement of men and women laureates scientific achievements. This leads to an observation that the quality of research work by women is equivalent to that of men, but women's participation and visibility in these fields are much less prominent.

There are number of papers that focus on the gender gap in publication productivity which state that there is a progress with regard to publication but women are still far from reaching parity across fields globally. Also, there are fields like mathematics, physics, computer science which might take longer time (maybe beyond this century) compared to other fields to reach equilibrium in terms of publications[Aguinis et al, 2018]. So far we have not reached a specific pattern of this asymmetry on the publication productivity in terms of fields, period, or geographical location. So it will be interesting to systematically analyze and arrive at the directions to be taken forward.

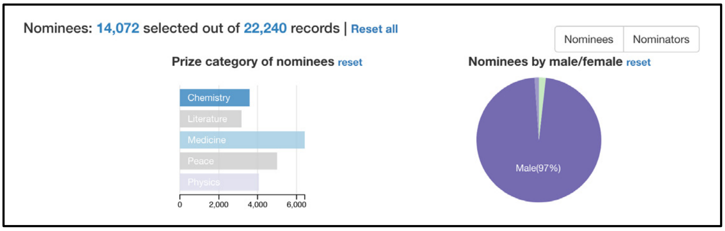
Table 3: T-test: Scientometric data

Scientometric Data	t_{calc}	$t_{critical}$
h-index	0.04	1.96
Number of citations by authors	0.01	1.96
Number of total documents by awardees	0.01	1.96



Nominations vs Laureates

Currently, data regarding nominations of candidates is available for the period 1901 - 1963 for Physics and Chemistry whereas for Medicine or Physiology, data is available until 1953. This is in compliance to the statutes of the Nobel Foundation i.e. only material older than 50 years can go public. From this data, it is clear that in all 3 fields, women were nominated at only 2% overall [Fig. 3]. This includes multiple nominators nominating one candidate. Considering the data for actual number of women nominated and winning the Nobel Prize, it is very encouraging i.e. 6 awards received out of 24 women nominated for the above said period. So, this leads to one important observation that the glass ceiling prevails at the initial of the process i.e. nomination itself. Hence, this has to be addressed appropriately by the awarding bodies without delay to improve the gender diversity in these awards.



Nominees: 14,072 selected out of 22,240 records | Reset all

Figure 3 Gender distribution of nominees in the fields of Chemistry, Physics & Medicine

Conclusion

The concern over very low representation of women in awards like the Nobel prize and Abel Prizes, Fields Medal and Turing awards, particularly in STEM fields, is often addressed of late. Few known and simple facts are educational empowerment, women's workforce, space for professional enhancement and the visibility of successful researchers. If we look at women receiving these awards considering the age, demographical location, scientometric analysis and quality of work it appears that there is no substantial difference with regard to their male counterpart. On the other hand, while looking at nominations, we find that women's representation is very low. In spite of governments and associated organizations working on schemes and frameworks to give due weight to women so that their knowledge, and their research output are suitably utilized, it appears there is still a huge scope to study important factors for narrowing the gap in such elite recognitions.

References

Aguinis, Herman, Young Hun Ji, and Harry Joo (2018), "Gender Productivity Gap among Star Performers in STEM and Other Scientific Fields." *Journal of Applied Psychology* 103 (12): 1283–1306.

Angayar Pavanasam, Aparna Basu (2019), Asymmetry:Women laureates in STEM through Scientometric Analysis, *Proceedings of Indo-German Workshop on Information Retrieval, Informetrics & Scientometrics*, Germany.

Benjamin J Fones and Bruce A Weinberg (2011), "Age dynamics in scientific creativity", *PNAS*, 108 (47), 18910-18914.

David Beede, Tiffany Julian, David Langdon, George McKittrick, Beethika Khan, Mark Doms (2011), "Women in STEM: A gender gap to Innovation, U.S. Department of Commerce, Economics and Statistics Administration, *ESA Issue Brief #04-11*.

Fulvio Naldi, Daniela Luzi, Adriana Valente, and Ilaria Vannini Parenti (2004), Scientific and Technological Performance by Gender, *Handbook of Quantitative Science and Technology Research*, Kluwer Academic Publishers, 299-314.

https://www.nobelprize.org/

https://www.abelprize.no/c53671/artikkel/vis.html?tid=53702

https://www.mathunion.org/imu-awards/fields-medal

https://amturing.acm.org/alphabetical.cfm

https://www.natureindex.com/news-blog/the-nobel-gender-gap-is-worse-than-you-think

Research performance and scholarly communication profile of competitive research funding: the case of Academy of Finland

Janne Pölönen[1] and Otto Auranen[2]

[1] *janne.polonen@tsv.fi*
Federation of Finnish Learned Societies, Snellmaninkatu 13, 00170 Helsinki (Finland)

[2] *otto.auranen@aka.fi*
Academy of Finland, Hakaniemenranta 6, 00531 Helsinki (Finland)

Abstract

The Academy of Finland (AKA), Finland's major public research funding agency, uses a Web of Science (WoS) based bibliometric indicator to assess the performance of research it has funded. In this paper, we use an alternative methodology to compare 1) the research performance and 2) scholarly communication profile of AKA funded research to the 13 Finnish universities' entire output across the major fields of science using the national data of 142,742 publications (years 2015-2018) registered in the VIRTA Publication Information Service. Research performance is analyzed using the Finnish Publication Forum (JUFO) rating of publication channels. Our results show that AKA-funded research is published on higher JUFO level journals and book publishers than Finnish universities' entire output, and that AKA-funded research is more focused than the universities on using peer-reviewed publications, articles published in journals, English language, and open access publishing. JUFO levels and WoS-based indicator produce consonant results in regard to the performance of AKA-funded research across all fields. The national publication data also provides added value by enabling assessment of the broader publication profile of AKA-funded research, including science communication, bibliodiversity, multilingualism, collaboration and open access.

Introduction

Like many research funders, the Academy of Finland (AKA) as a major Finnish public research funding agency is interested in the effects of its funding. As a partial measure for its goal of promoting high quality, renewal and impact of research, AKA uses a bibliometric indicator to compare the citation impact of AKA-funded publications to the citation impact of the entire Finland. Indicator has been in use since 2016 and the main elements of method have been described by Auranen & Leino (2019). The indicator is called the Top10 index, which describes the unit of analysis' relative share of the 10% most cited publications in the world, world average being 1 (Academy of Finland, 2020a). Web of Science (WoS) is used as a data source for calculating the Top10 index. Publications from two AKA's funding instruments are included in the calculation of the indicator: Academy Projects and Academy Research Fellows, which represent bottom-up funding instruments in AKA's funding portfolio (Academy of Finland, 2020b), and of Fellows: Academy of Finland, 2020c). Results show that AKA-funded publications have a higher Top10 index than publications from Finland on average (1.29 vs 1.09 in 2020). However, Top10 index for Academy Research Fellows is usually higher than for Academy Projects (1.38 vs 1.26 in 2020).

Web of Science allows international bench-marking, however the disadvantage is the narrow focus on peer-reviewed international journal articles. Comparisons with the comprehensive institutional CRIS data, which in some countries - including Finland - has been integrated at the national level, have shown that WoS and Scopus coverage is seriously lacking in the social sciences and the humanities (SSH) (van Leeuwen et al., 2016; Aksnes & Sivertsen, 2019; Sivertsen, 2019; Pölönen et al., 2020b). This is because in the SSH fields, local publication languages and books play an important role in the scholarly communication (Kulczycki et al., 2020; Kulczycki et al., 2018; Engels et al. 2018). Due to lack of WoS coverage, AKA's bibliometric indicator misses a large share of scientific publishing particularly in the SSH, as 41 % of the peer-reviewed articles reported from the social sciences and 79 % from the humanities are not included in the indicator (Auranen & Leino 2019). In addition, especially in the SSH but also in other fields, dissemination of research knowledge within and beyond academia involves a broad range of publications that are not peer-reviewed (Hicks, 2004). Also these types of publications, which are highly relevant for the societal impact of research, are excluded from the WoS-based analyses but they are covered in the CRIS data. In Finland, peer-reviewed and not-peer-reviewed publications are comprehensively covered in the national VIRTA publication information service, which integrates the publication metadata from the local CRIS of 13 Finnish universities (Pölönen, 2018).

One challenge in using the national publication data for assessing research performance, as opposed to WoS or Scopus, is the lack of citation data. In several countries, performance-based research funding systems (PRFS) use a comprehensive list of peer-reviewed publication channels as a quality-index (Hicks, 2012; Sivertsen, 2016; Pölönen et al, 2020a). In Finland, Publication Forum classification (in short, JUFO) has been developed since 2010 to support the Ministry of Education and Culture's PRFS for allocating part of core-funding annually to universities. The main rationale is to reward universities not only based on quantity but also quality of output, namely publishing in channels that are valued by the scientific community, are demanding in terms of peer reviews, and reach the widest critical expert audience. Several Finnish universities have also used local CRIS data and JUFO levels, in addition to bibliometric citation analyses based on WoS or Scopus data, to inform expert-panels conducting institutional research assessments (Wang et al., 2014; Pölönen et al, 2021). So far, the national VIRTA data or JUFO classifications have not been used to assess the performance or broader scholarly communication practices of the AKA funded research.

In JUFO classification (Publication Forum, 2021), domestic and foreign peer-reviewed publication channels (journals and book publishers) are classified to four level categories (1=basic, 2=leading, 3=top, 0=other) by Finnish experts in the field (Auranen & Pölönen, 2012; Pölönen et al., 2018). The evaluation of channels is entrusted to 250 experts in 23 field-specific panels, who represent the Finnish research community. The experts' main tasks are 1) to identify reliable peer-reviewed channels, and 2) indicate the leading channels of their field in terms of average quality, impact and prestige. JUFO evaluation is informed but not constrained by citation-based journal metrics, such as the Journal Impact Factor (JIF) or Source Normalized Impact per Paper (SNIP). In the SSH fields, also national language journals and publishers without impact factors are included among the leading channels (Pölönen et al., 2021).

Research questions, data and methods

According to the international initiatives for responsible metrics, evaluation should take into consideration the disciplinary diversity and plurality of research outputs (https://sfdora.org; Hicks et al., 2015; Wilsdon et al., 2015). Also the Responsible Research and Innovation (RRI) and Open Science (OS) policies have called for a broader-based evaluation of research taking into account societal interaction and impact. In this study, we compare the research performance and scholarly communication profiles of the Finnish universities and the AKA-funded research based on comprehensive national publication data and the JUFO classification. We pose the following two research questions:

1. How does AKA-funded research at Finnish universities perform in publishing compared with entire research activity at Finnish universities? We use the share of peer-reviewed outputs published in journals, conferences and book publishers in JUFO levels 2 ("leading") and 3 ("top") as the indicator of publication performance. We also compare the results based on JUFO levels with those based on the WoS-based Top10 index.
2. What is the scholarly communication profile of AKA-funded research at Finnish universities in comparison with the entire publication output of Finnish universities?

We compare the scholarly communication profiles according to the following indicators derived from VIRTA data across the main fields of science:

1. Science communication: share of not-peer-reviewed publications aimed at academic, professional and general audiences.
2. Bibliodiversity: share of peer-reviewed book publications (chapters, monographs and edited volumes), conference articles and conference articles.
3. Multilingualism: share of peer-reviewed publications in languages other than English (Finnish, Swedish and other languages).
4. Open access: share of peer-reviewed open access publications, including gold, hybrid and green OA.
5. Research collaboration: share of peer-reviewed publications with co-authors, with co-authors from more than one Finnish university, and international co-authors.

To investigate these research questions, we created a dataset based on three sources:

1. VIRTA publication data, consisting of 158,029 publications (publication year 2015-18) the 13 Finnish universities have validated and reported annually to the Ministry of Education and Culture. The Ministry used this data to allocate 13% of core funding annually to universities in the performance-based research funding system . Publications with authors affiliated with more than one Finnish university figure in the data as duplicates.
2. AKA publication data, consisting of 7,971 publications (publication year 2015-18), which the PIs of the Academy Projects and Academy Research Fellows from the call years 2011-2013 have reported as outputs. Publication reported by more than one project or fellow figure in the data as duplicates. Since 2017, AKA-funded research outputs have been reported using VIRTA as one of the information sources.
3. JUFO publication channel classification, consisting of 31,597 journals and series, conferences and book publishers evaluated and rated according to quality, impact and prestige by national expert-panels. Information about the channel identified by unique JUFOID and the JUFO level is included in the VIRTA data for all peer-reviewed publications.

We matched AKA and VIRTA publications, and indicated in the VIRTA data which publications by the Finnish universities have been produced with funding for AKA Projects and Fellows (Table 1). We also deduplicated the VIRTA data to arrive at full-counts at national level. Our final dataset consists of 142,742 Finnish universities' publications, of which 6,143 (4%) are an AKA-funded subset. Dataset comprises only publications where at least one of the authors is affiliated at a university. It should be noted that while the Academy of Finland grants funding to researchers with various affiliations, appr. 80% of the funding is granted to researchers at universities (see Academy of Finland, 2020d).

The identification of peer-review status, target audience, publication type, language, open access status, number of authors and international co-authorship of publications in VIRTA is based on researchers' self-reports and/or validation by the data-collection personnel at the universities. Over two-thirds (69%) of all Finnish Universities' publications reported in VIRTA are peer-reviewed scientific publications, including articles in journals, conferences and books, as well as monographs and edited volumes (Table 1). Almost one-third of the outputs are not peer-reviewed publications for the academic, professional and general audiences. Of 6,143 publications reported as outputs of AKA-funded research, the vast majority (89%) are peer-reviewed.

Table 1. Finnish universities' publications and Academy of Finland (AKA) funded publications 2015-2018 by scholarly status

SCHOLARLY STATUS	ALL PUBLICATIONS		AKA-FUNDED	
	Number	Share	Number	Share
Scientific peer-reviewed	98472	69.0 %	5478	89.2 %
Not peer-reviewed	44270	31.0 %	665	10.8 %
All	142742	100 %	6143	100 %

Of the AKA-funded publications, 4,359 are reported as Projects' outputs and 1,896 as Fellows' outputs (Table 2). This includes 112 publications that have been reported as outputs of both AKA-funded Projects and Fellows. All publication outputs are assigned to OECD FOS main fields based on the field-classification in the VIRTA data based on researchers' self-reports. The share of Natural sciences and Engineering publications is larger, and that of SSH publications smaller, in the AKA-funded output compared to the universities' entire output.

Table 2. Finnish universities' publications and Academy of Finland (AKA) funded Projects' and Fellows' publications 2015-2018 by main field of science

MAIN FIELD	ALL PUBLICATIONS		AKA-FUNDED		AKA PROJECTS		AKA FELLOWS	
	Number	Share	Number	Share	Number	Share	Number	Share
1 Natural sciences	34565	24 %	2449	40 %	1684	39 %	809	43 %
2 Engineering	15464	11 %	784	13 %	429	10 %	370	20 %
3 Medicine	27387	19 %	1144	19 %	942	22 %	217	11 %
4 Agriculture	3077	2 %	89	1 %	74	2 %	16	1 %
5 Social sciences	37259	26 %	1051	17 %	752	17 %	315	17 %
6 Humanities	24990	18 %	626	10 %	478	11 %	170	9 %
All fields	142742	100 %	6143	100 %	4359	100 %	1896	100 %

Results

Publication performance of AKA-funded research compared to Finnish universities' entire peer-reviewed publication output in 2015-2018

In this part, we limit our analysis to 98,472 peer-reviewed publications, of which 5,478 (5.6%) are AKA-funded research outputs, including 3,892 (4%) outputs related to AKA Projects and 1,681 (1.7%) to AKA Fellows. Overall, both the universities' peer-reviewed output and the AKA-funded research is published in channels in all JUFO level categories from 0 to 3 (Table 3). The AKA-funded research is, however, more strongly concentrated to channels the national expert-panels have rated as JUFO level 2 "leading" and 3 "top" channels. The share of JUFO level 2 and 3 publications is 45% for the AKA-funded research, compared to 32 % for the Finnish universities in general (Figure 1). Also, a larger share of publications by AKA-funded Fellows (48 %) is on JUFO levels 2 and 3 than publications by AKA-funded Projects (43 %). Our analysis also shows that AKA-funded research outperformed Finnish universities in all fields, and in case of each university's peer-reviewed output (Figure 2). These results are consonant with the WoS-based bibliometric analysis, as the AKA-funded research has a higher Top10 index than the Finnish research in general, and also the research by AKA Fellows has a higher Top10 index than AKA projects.

Table 3. Research performance: share of peer-reviewed publications of Finnish universities and Academy of Finland (AKA) funded research according to JUFO level of the publication channel

JUFO-LEVEL	ALL PUBLICATIONS		AKA-FUNDED		AKA PROJECTS		AKA FELLOWS	
	Number	Share	Number	Share	Number	Share	Number	Share
3	9372	9.5 %	739	13.5 %	502	12.9 %	250	14.9 %
2	22425	22.8 %	1699	31.0 %	1170	30.1 %	564	33.5 %
1	57194	58.1 %	2766	50.5 %	2038	52.4 %	770	45.8 %
0	9481	9.6 %	274	5.0 %	182	4.7 %	98	5.8 %
All JUFO levels	98473	100 %	5478	100 %	3893	100 %	1681	100 %

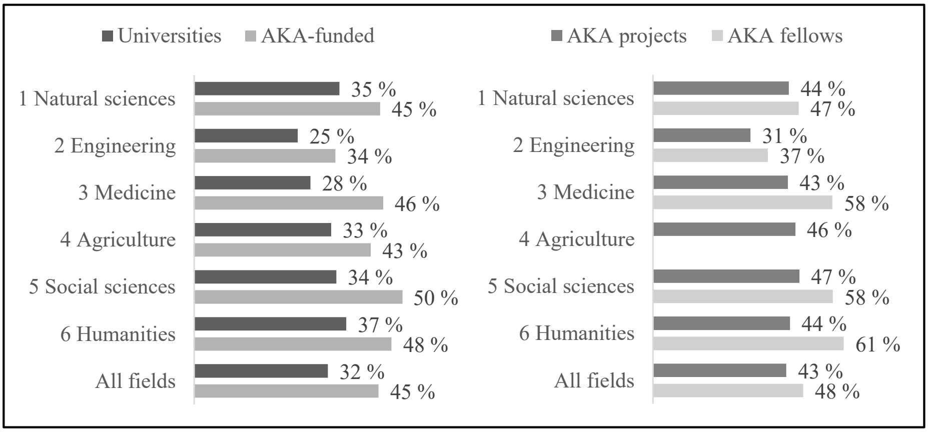


Figure 1. Share of JUFO-level 2 and 3 channels of all peer-reviewed publications of Finnish universities and AKA-funded projects and fellows by the main field of science.

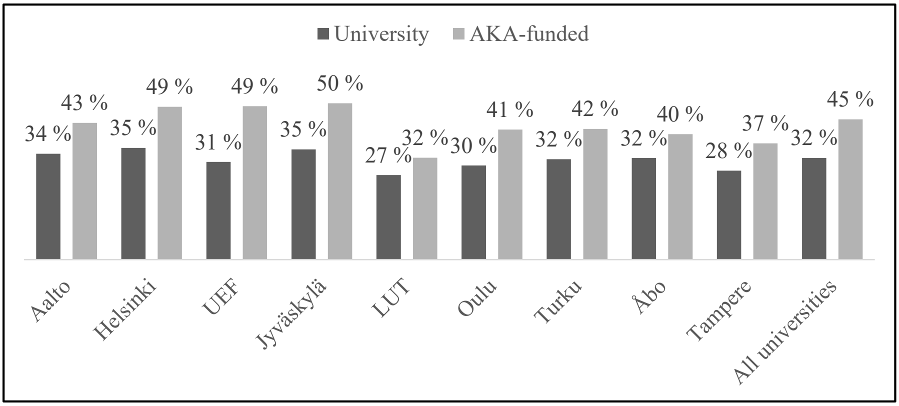


Figure 2. Share of JUFO-level 2 and 3 channels of all peer-reviewed publications of 9 Finnish universities and AKA-funded projects and fellows (Lapland, Hanken, UNIArts and Vaasa are excluded because the number of AKA-funded outputs is less than 50).

Publication profiles of AKA-funded research compared to Finnish universities' entire publication output in 2015-2018

1. Science communication. According to our analysis, both Finnish universities and AKA-funded research (Fellows and Projects) use a broad range of publications, mostly peer-reviewed publications but also non-scholarly publications aimed at academic, professional and general audiences (Table 4). The AKA-funded research is, however, much more focused on peer-reviewed scholarly communication within academia. The share of not-peer-reviewed publications of the total output is 11% for the AKA-funded research compared to 31% in the case of Finnish universities. The share of not-peer-reviewed publications is much larger in the SSH fields in case of both AKA-funded research and Finnish universities' output (Figure 3).

2. Bibliodiversity. Both Finnish universities and AKA-funded research use peer-reviewed journal, conference and book publications for scholarly communication (Table 4). Nevertheless, the AKA-funded research is more focused on using peer-reviewed journal articles. The share of articles in conferences and books, as well as monographs, of the total peer-reviewed output is 22% for the AKA-funded research compared to 29% in the case of Finnish universities. Both AKA-funded research and Finnish universities' output shows traditional disciplinary differences in use of different publication types, especially in engineering (conferences) and the SSH (book publications). Interestingly, in the case of humanities the share of journal articles is slightly smaller for AKA-funded research than Finnish universities in general (Figure 3).

3. Multilingualism. Both Finnish universities and AKA-funded research use multiple languages in peer-reviewed scholarly communication, including English, Finnish and Swedish (Finland's two national languages), as well as other languages (Table 4). The AKA-funded research is, however, much more focused on English language publications. The share of peer-reviewed publications in languages other than English is 4,5 % for the AKA-funded research compared to 11 % in the case of Finnish universities' entire peer-reviewed output. Despite the strong focus on English language publishing, the share of publications in languages other than English is much larger in the SSH fields in case of both AKA-funded and Finnish universities' output (Figure 3).

4. Open Access. Both Finnish universities and AKA-funded research use different routes, including gold, hybrid and green OA, to enable open access to peer-reviewed publication outputs (Table 4). Overall, the share of Open Access publications is larger in case of the AKA-funded research. While the share of outputs in gold OA channels is almost the same, AKA-funded research has a larger share of hybrid and green (self-archived) OA outputs. Overall, the share of Open Access peer-reviewed output is 53% for the AKA-funded research compared to 48% in the case of Finnish universities. There are some differences between fields in the overall Open Access share, however the Open Access advantage of AKA-funded research is stronger in medicine, agriculture and social sciences, and non-existent or modest in case of natural sciences, engineering and humanities (Figure 3).

5. Collaboration. According to our analysis, both Finnish universities and AKA-funded research produce the vast majority of peer-reviewed publications in collaboration between two or more authors, who are often affiliated with other Finnish or foreign universities (Table 4). The AKA-funded research is, however, more focused on research collaboration. The share of co-authored publications of the total peer-reviewed output is 91 % for the AKA-funded research compared to 82 % in the case of Finnish universities. Also the share outputs produced in Finnish inter-university collaboration and international collaboration is larger for the AKA-funded research (23% and 50%, respectively) compared to Finnish universities in general (12% and

46%, respectively). Single-authorship is most common in the SSH fields, so the increased collaboration related to AKA-funded research is visible especially in these fields (Figure 3). Perhaps surprisingly, the share of internationally co-authored publications is smaller for AKA-funded research in all fields except medicine.

Table 4. Scholarly communication profiles of Finnish universities and the AKA-funded research

SCHOLARLY COMMUNICATION PROFILE	**ALL PUBLICATIONS**		**AKA-FUNDED**	
	Number	**Share**	**Number**	**Share**
1. Science communication				
● Peer-reviewed	98472	69.0 %	5478	89.2 %
● Academic	14928	10.5 %	372	6.1 %
● Professional	18240	12.8 %	163	2.7 %
● General	11101	7.8 %	130	2.1 %
● All	142742	100 %	6143	100 %
2. Bibliodiversity				
● Journal articles	70050	71.1 %	4265	77.9 %
● Conference articles	13250	13.5 %	679	12.4 %
● Book articles	14112	14.3 %	491	9.0 %
● Monographs	1061	1.1 %	43	0.8 %
● All	98472	100 %	5478	100 %
3. Multilingualism				
● English	87323	88.7 %	5233	95.5 %
● Finnish	8835	9.0 %	209	3.8 %
● Swedish	882	0.9 %	6	0.1 %
● Other	1433	1.5 %	30	0.5 %
● All	98472	100 %	5478	100 %
4. Open Access				
● Gold OA	16458	22.0 %	902	21.7 %
● Hybrid OA	6057	8.1 %	439	10.5 %
● Green OA	13358	17.9 %	848	20.4 %
● Closed	38891	52.0 %	1975	47.4 %
● All	98472	100 %	5478	100 %
5. Collaboration				
● Single-authored	17566	17.8 %	500	9.1 %
● Co-authored	80907	82.2 %	4978	90.9 %
● - Inter-university co-authors	11778	12.0 %	1293	23.6 %
● - International co-authors	45186	45.9 %	2729	49.8 %
● All	98472	100 %	5478	100 %

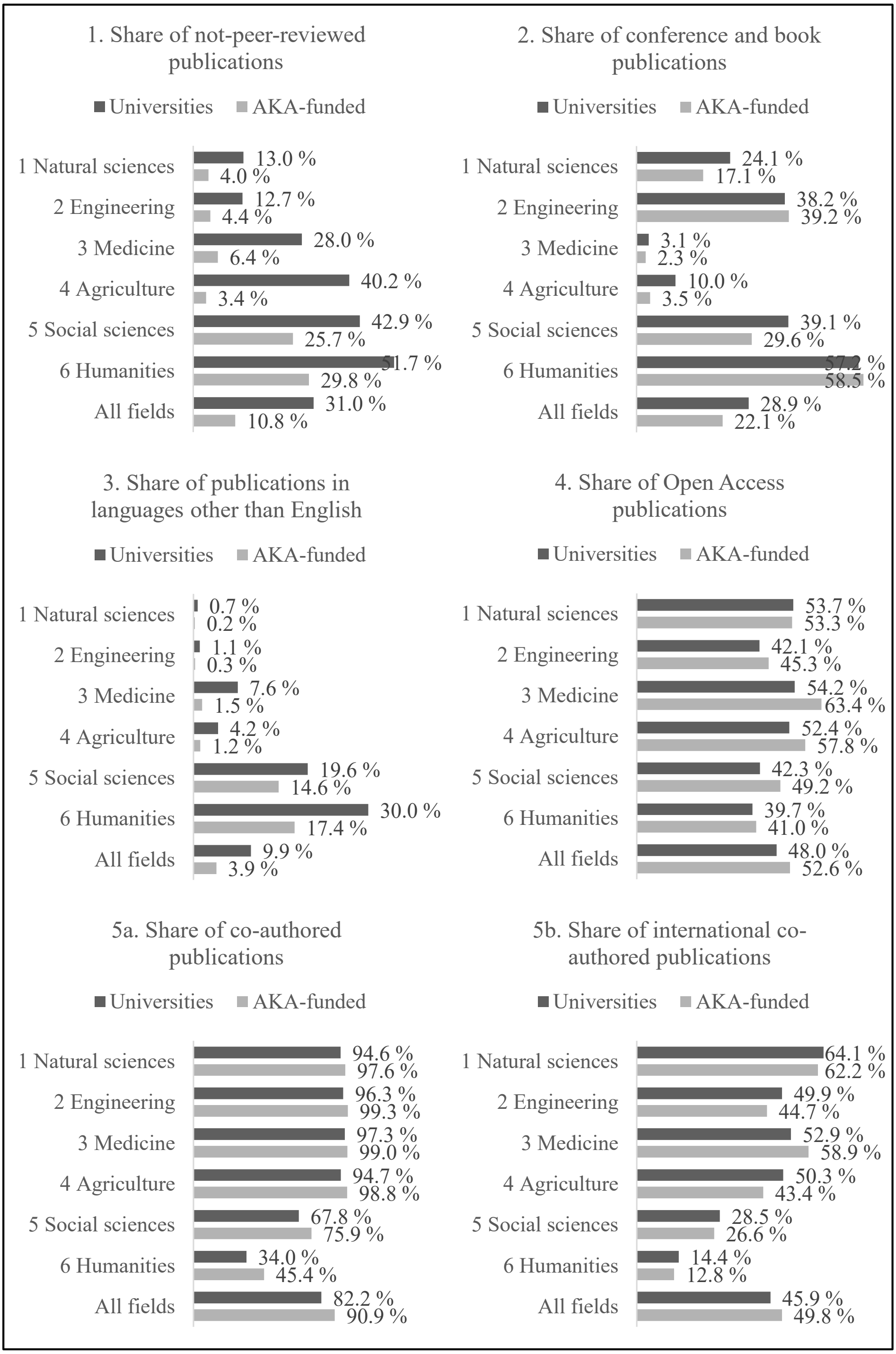


Figure 3. Scholarly communication profiles of Finnish universities and AKA-funded Projects and Fellows by main fields of science.

Discussion and conclusions

In this study we first compared the research performance of AKA-funded research and research conducted in the Finnish universities in general. Instead of bibliometric citation analysis based on WoS data, we measured research performance by using comprehensive national publication data including all publication types and languages, and the Finnish national expert-based JUFO classification of journals and book publishers as a quality-index. Indicator based on JUFO levels and the WoS based Top10 index produced consonant results, as the AKA-funded research outperforms the baseline (Finland or Finnish universities) both in the citation impact as well as in the share of peer-reviewed outputs published in journals, conferences and book publishers the JUFO expert-panels have nominated as "leading" publication channels (levels 2 and 3) (Figure 4). Similarly, AKA Fellows outperformed AKA Projects based on both indicators. The same was observed in the case of individual universities as well. JUFO-based indicator has the advantage of taking into account the disciplinary diversity and plurality of research outputs.

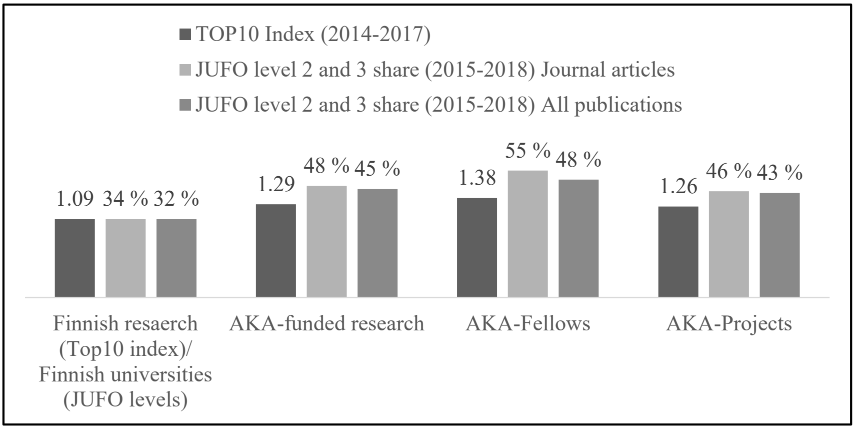


Figure 4. Comparison of Top10-index and share of JUFO levels 2 and 3 outputs as research performance indicators for Finnish research/Finnish universities, AKA-funded research, AKA-funded Fellows, and AKA-funded Projects. Both indicators have been calculated including all peer-reviewed publication types: journal, conference and book articles, as well as monographs.

The national publication data also provides added value by enabling assessment of the scholarly communication profile. We looked at publication profiles in the second part of this study, and discovered that both the Finnish universities and AKA-funded research take care of the responsibilities of science communication by disseminating research knowledge in not-peer-reviewed publications aimed at the academic, professional and general audiences. Both AKA-funded research and universities also show considerable bibliodiversity and multilingualism in peer-reviewed scholarly communication, and both show the same traditional field specific differences in target audiences, publication types, languages, and collaboration. Nevertheless, the AKA-funded research is also more focused than the universities in general on using peer-reviewed publications, articles published in journals, and English as publication language. Also, a larger share of AKA funded research is Open Access.

The Academy of Finland employs international peer-review to select most promising research for funding, with a goal of promoting high quality, renewal and impact of research. Furthermore, the AKA funding instruments included in our analysis are bottom-up funding opportunities with an emphasis on high quality of (basic) research, international collaboration and established position in the international scientific community. Funding for AKA Projects and Fellows is also rather selective, with success rates hovering between 12 and 15 % in recent years (see Table 1 at Academy of Finland, 2020e). It is expected that AKA-funded research would have a strong scientific impact internationally, as indeed shown by Top10 index. Strong emphasis on international impact and application success rates may also explain our findings that AKA-funded research is published in leading journals and book publishers, and with relatively strong preference for English as a publication language. AKA is interested in the societal impact of the research conducted by Projects and Fellows; information about this is requested both in the application and reporting phase. However, our results show that in publishing, science communication is not their priority. AKA funding criteria and policies do not privilege journal publishing as such over conference and book publishing, and indeed our findings suggest that also AKA-funded research follows traditional disciplinary patterns in scholarly communication and collaboration. As a cOAlition S member, AKA is strongly committed to Open Access to research published with its funding (see Academy of Finland, 2020f), and this policy readily explains the larger OA share of AKA-funded research.

One of the limitations of our study is that we were not able to analytically examine the influence of disciplinary variations in comparison of AKA Top10 index and share of JUFO levels 2 and 3 outputs. The main obstacle is the limited number of outputs covered by WoS in some SSH fields, as well as the small number of AKA-funded outputs in VIRTA when differentiated according to the main field. Another challenge for this comparison is due to specific disciplinary grouping used in AKA's WoS-based bibliometric analyses that is difficult to match to disciplinary classification used in VIRTA publication data.

Another possible limitation relates to reporting of outputs by researchers to institutional CRIS, from which VIRTA data is integrated, as well as reporting of outputs of AKA-funded research by the principal investigators (PI). In both cases, reporting may be less comprehensive in case of the other than peer-reviewed outputs, and there may be field-specific differences. In general, universities have a financial incentive to report outputs comprehensively, including publications for professional and general audiences as these are included among the Ministry's funding criteria since 2015. It is not clear, however, if PIs of the AKA-funded projects comprehensively report these types of publications, or if they tend to overreport the output of their projects. In addition, PIs may report only outputs published by the time they submit the final report. Reporting usually takes place several months after the termination of the project, but due to publication delays we have not been able to identify all AKA-funded outputs.

Future analyses could extend comparison to other funding instruments, of which funding programmes of the Strategic research council (STN) - hosted by AKA - would be of particular interest. This is because STN funding programmes emphasise the societal impact and interaction, which might be expected to result in somewhat different scholarly communication profiles compared to AKA funding instruments.

In all, we conclude that national VIRTA data and JUFO levels complement WoS based bibliometric analyses of research performance with a comprehensive coverage of publication output including all fields, publication types and languages. They offer a relevant information source for a responsible macro level assessment and monitoring of publication activity in several contexts in Finland, for example research assessment of organizations and research fields, or analyses of research funded via competitive mechanisms, such as the Academy of Finland funding. Currently institutional and national CRIS data cover the diversity of research outputs, however they do not support international comparisons of research performance and scholarly communication profiles. To enhance international comparisons, we suggest international integration of institutional and national publication data (Sivertsen, 2019; Puuska et al, 2020).

References

Academy of Finland (2020a). *Concepts related to bibliometric analyses*. https://wiki.eduuni.fi/pages/viewpage.action?pageId=138151334#Bibliometrisiinanalyyseihinliittyvi%C3%A4k%C3%A4sitteit%C3%A4-Conceptsrelatedtobibliometricanalyses

Academy of Finland (2020b). *Funding for research teams*. https://www.aka.fi/en/research-funding/funding-opportunities-at-a-glance/funding-for-research-teams/

Academy of Finland (2020c). *Funding for individual researchers*. https://www.aka.fi/en/research-funding/funding-opportunities-at-a-glance/funding-for-individual-researchers/

Academy of Finland (2020d). *Who gets the funding*. https://www.aka.fi/en/about-us/what-we-do/what-we-are/who-gets-the-funding/

Academy of Finland (2020e). *Application and funding statistics*. https://www.aka.fi/en/about-us/data-and-analysis/application-and-funding-statistics/

Academy of Finland (2020f). *Open science: open access publishing and open data*. https://www.aka.fi/en/research-funding/responsible-science/open-science/

Aksnes, D. W. & Sivertsen, G. (2019). A Criteria-based Assessment of the Coverage of Scopus and Web of Science. *Journal of Data and Information Science*, 4, 1-21. https://doi.org/10.2478/jdis-2019-0001

Auranen, O. & Leino, Y. (2019). Bibliometric indicator to assess the effectiveness of competitive research funding, 24th Nordic workshop on bibliometrics and research policy, Reykjavik.

Auranen, O. & Pölönen, J. (2012). *Classification of scientific publication channels: Final report of the Publication Forum project (2010–2012)*. Helsinki: Federation of Finnish Learned Societies. Http://www.julkaisufoorumi.fi/sites/julkaisufoorumi.fi/files/publication_forum_project_final_report_0.pdf.

Engels, T.C.E., Starčič, A., Kulczycki, E., Pölönen, J. & Sivertsen, G. (2018). Are book publications disappearing from scholarly communication in the social sciences and humanities? *Aslib Journal of Information Management*, 70, 592-607. https://doi.org/10.1108/AJIM-05-2018-0127

Hicks, D. (2004). The four literatures of social science. In Moed H. (Ed). *Handbook of quantitative science and technology research* (pp. 473-496). Dordrecht: Kluwer Academic.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41, 251-261. Https://doi.org/10.1016/j.respol.2011.09.007

Hicks, D., Wouters, P. F., Waltman, L., de Rijcke, S., and Rafols, I. (2015). The Leiden Manifesto for research metrics: use these 10 principles to guide research evaluation. *Nature*, 520, 429-431. Https://doi.org/10.1038/520429a.

Kulczycki, E., Engels, T.C.E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., Nowotniak, R., Petr, M., Sivertsen, G., Istenič Starčič, A. & Zuccala, A. (2018). Publication patterns in the social sciences and humanities: The evidence from eight European countries. *Scientometrics*, 116, 463-486. https://doi.org/10.1007/s11192-018-2711-0

Kulczycki, E., Guns, R., Pölönen, J., Engels, T. C. E., Rozkosz, E. A., Zuccala, A. A., Bruun, K., Eskola, O., Starčič, A. I., Petr, M., & Sivertsen, G. (2020). Multilingual Publishing in the Social Sciences and Humanities: A Seven-Country European Study. *Journal of the Association for Information Science and Technology*, 71, 1371-1385. https://doi.org/10.1002/asi.24336

Publication Forum (2021). *Publication Forum*. https://www.julkaisufoorumi.fi/en

Puuska, H.-M., Nikkanen, J., Engels, T., Guns, R., Ivanović, D. & Pölönen, J. (2020). Integration of national publication databases — towards a high-quality and comprehensive information base on scholarly publications in Europe. *Proceedings of the International Conference on ICT enhanced Social Sciences and Humanities 2020*. https://doi.org/10.1051/itmconf/20203302001

Pölönen, J. (2018). Applications of, and Experiences with, the Norwegian Model in Finland. *Journal of Data and Information Science*, 3, 31-44. https://doi.org/10.2478/jdis-2018-0019

Pölönen, J., Auranen, O., Engels, T. & Kulczycki, E. (2018). Taking national language publications into account: the case of the Finnish performance-based research funding system. *STI 2018 Conference Proceedings* (pp. 204-211). Leiden: University of Leiden. Https://openaccess.leidenuniv.nl/bitstream/handle/1887/65223/STI2018_paper_43.pdf?sequence=1

Pölönen, J., Guns, R., Kulczycki, E., Sivertsen, G., & Engels, T. C. E. (2020a). National Lists of Scholarly Publication Channels: An Overview and Recommendations for Their Construction and

Maintenance. *Journal of Data and Information Science* (published online ahead of print). doi: https://doi.org/10.2478/jdis-2021-0004

Pölönen, J., Laakso, M., Guns, R., Kulczycki, E. & Sivertsen, G. (2020b). Open access at the national level: A comprehensive analysis of publications by Finnish researchers. *Quantitative Science Studies*, 1, 1396-1428. Https://doi.org/10.1162/qss_a_00084

Pölönen, J., Pylvänäinen, E., Aspara, J., Puuska, H.-M. & Rinne, R. (2021). *Publication Forum 2010-2020: Self-evaluation report of the Finnish quality classification system of peer-reviewed publication channels*. Helsinki: Federation of Finnish Learned Societies. https://julkaisufoorumi.fi/sites/default/files/2021-03/Publication%20Forum%20self-evaluation%20report%202021_0.pdf

Sivertsen, G. (2016). *Publication-based funding: The Norwegian model*. In M. Ochsner (Eds.), Research Assessment in the Humanities: Towards Criteria and Procedures (pp. 71–90). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-29016-4_7

Sivertsen, G. (2019). *Developing Current Research Information Systems (CRIS) as data sources for studies of research*. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), Springer Handbook of Science and Technology Indicators (pp. 667–683). Cham: Springer.

van Leeuwen, T. N., van Wijk, E., & Wouters, P. F. (2016). Bibliometric analysis of output and impact based on CRIS data: A case study on the registered output of a Dutch university. *Scientometrics*, 106, 1-16. Https://doi.org/10.1007/s11192-015-1788-y

Wang, L., Vuolanto, P. & Muhonen, R. (2014). *Bibliometrics in the research assessment exercise reports of Finnish universities and the relevant international perspectives*. Tampere: University of Tampere. Http://tampub.uta.fi/handle/10024/96266.

Research performance indicators and management decision making: Using Collab-CNCI to understand institutional impact

Ross W. K. Potter[1], Martin Szomszor[2] and Jonathan Adams[3,4]

[1] *ross.potter@clarivate.com,* [2] *martin.szomszor@clarivate.com,* [3] *jonathan.adams@clarivate.com*

[1,2,3] Institute for Scientific Information, Clarivate, 160 Blackfriars Road, London, SE1 8EZ (United Kingdom)

[4] Policy Institute, King's College London, 22 Kingsway, London, WC2B 6LE (United Kingdom)

Abstract

Normalized citation metrics are used to analyze, manage, and fund research. The standard category normalized citation index (CNCI) metric, however, obfuscates individual entities and the role of collaboration. Given the increasing importance of collaboration, normalized metrics require additional context for decision making. Here, we compare different CNCI calculation methods for 30 institutions (covering six geo-political regions): the conventional CNCI method, Collab-CNCI (which considers collaboration), and Fractional CNCI (assigning credit). Analyzing article data from 2009-2018 highlights the steady global increase in international collaboration. Normalizing by collaboration type, Collab-CNCI suppresses the impact of multi-lateral-authored articles on CNCI mean for all institutions. Consequently, Collab-CNCI and Fractional CNCI produce comparable results but, crucially, Collab-CNCI does not assign potentially incorrect credit to institutions. Over the period, institutions in developing or newly established research economies generally show the greatest improvement in Collab-CNCI (e.g., Malaysia, China) demonstrating their research base growth; institutions in developed research economies (e.g., USA) generally show low or negative improvement, but still retain high Collab-CNCI values. The breakdown of domestic and international collaborative groups within Collab-CNCI allows managers to better contextualize institutional CNCI data and, therefore, make better informed funding decisions.

Introduction

Research activity data refer to inputs (facilities and money), activity (people and their projects), and outputs (papers and their citations); research evaluation and management require an analysis of the relationship between these. The 'value' of inputs and outputs in relation to activity differs between and within countries, between subjects, and varies temporally. Consequently, if a funder, researcher, or manager is to understand the performance of a research entity they need to adjust the raw values of both input and output data using appropriate benchmarks. Financial data are usually adjusted to Purchasing Power Parity (PPP) via internationally agreed models. Publication data are more problematic.

While the ultimate objective of research investment is the generation of knowledge that benefits wealth creation and quality of life, the proximate unit of research output is a published report of results and their significance. However, research publication rates are not constant between fields and the innate value of each output is unclear. Citations to a published work are conventionally seen as a reflection of the value that is placed on that work through its utility for later publications (Garfield, 1955). There is a growing body of work that has established that, for reasonably large samples, there is a sound relationship between citation counts (specifically, of journal articles) and peer judgments of quality in science, technology, and many social science fields, although this is less evident in the arts and humanities (Evidence, 2007; Waltman, 2016; Aksnes et al., 2019). Citations accumulate over time at a rate that is field dependent. Therefore, to create an index that is more informative for comparisons, raw counts are normalized by calculating the ratio between an observed count and the estimated global average for all papers published in the same year and subject category. This is referred to as Category Normalized Citation Impact (CNCI).

Average CNCI is widely used as a standard indicator for both national and institutional comparisons (Jappe, 2020) and is likely used by research managers within institutions as well. However, a single metric of this kind contains little information. It expresses the average ratio between the observed and expected counts for a set of documents but says nothing about the variance within the set, the proportion of high values, or the nature of the papers that contributed to the high values. It consequently has very limited value for management purposes since it has no explanatory power and suggests no action to improve or mitigate performance.

It has also become evident that collaborative papers are, on average, more frequently cited than single-author papers and that international collaboration is rising (Wagner & Leydesdorff, 2005; Wagner, 2008; Leydesdorff & Wagner, 2008) and associated with higher citation counts than purely national papers (Adams, 2012). There is sound reason to infer that international collaboration should be associated with higher impact research. The costs, as well as benefits, of collaboration are widely acknowledged (e.g., Katz & Martin, 1997; Bozeman et al., 2013) and it is reasonable to suppose that researchers normally engage in collaboration only to further such objectives as they cannot accomplish under their own control. Analysis shows that higher performing institutions typically have higher levels of international collaboration and that they tend to link to other institutions of a similar status; this is associated with exceptionally highly cited research outputs (Adams, 2013).

Some analysts suggest that the effect of collaboration, in raising average CNCI values, is in practice a distortion of the 'correct' citation impact index, if such a ground truth could be demonstrated (Waltman & van Eck, 2015). They suggest that both the credit for publishing and the credit for being cited should be apportioned fractionally among the authors, their institutions, and countries. The central benefit of carrying out such fractional counting is that country counts (e.g., of papers) can be summed to a correct world total (whereas it would otherwise be inflated by double counting) and that national average CNCI values can themselves be subsumed into a world average of 1.0 (whereas it is otherwise feasible for all countries to have a CNCI greater than the world average: Szomszor et al., 2021). This has significant arithmetic appeal and makes some derived calculations more satisfactory.

Because there is no source of truth as to the correct value of a CNCI indicator, however, it is unclear that fractional counting provides a result that delivers any management benefit: it is neither more precise nor more accurate than a full counting methodology. Furthermore, the assumption that a higher value of citation impact brought about by a collaboration should be 'corrected' and reduced by this procedure is a value laden judgment on the part of the analyst from which researchers might reasonably demur. Nonetheless, it remains true that collaborative papers do receive more citations, on average, and that highly collaborative papers (which may have thousands of authors) do appear to be of a different nature (by their authorship and their subsequent citation profile) from the bulk of research outputs (Aksnes, 2003).

The contribution of internationally collaborative papers to a higher average CNCI than a set of domestic papers on their own will not be apparent to an observer who sees only a reported net index value. Fractional counting will suppress the value (through apportionment) but still offer no management information on the result. The Institute for Scientific Information have therefore proposed an alternative approach to calculating citation impact to address the challenge of turning the data into management information (Potter et al., 2020). This variant index (Collab-CNCI) considers different levels of domestic and international authorship collaboration.

Domestic research (where publications carry author addresses from only one country) may belong to a single institution or be collaborative across several institutions, so this is a factor to consider since it may affect an institutional research strategy (Katz & Hicks, 1997). Internationally collaborative research may be bilateral (between two countries), trilateral (involving a third country) or quadrilateral-plus (involving four or more countries). Average CNCI rises from group to group with collaborative diversity (Adams & Gurney, 2018; Adams et al., 2019). The latter group could be subdivided further but they are relatively scarce in any sample. There is, however, an acknowledged argument that papers with large author counts

(~25-30 or more) should be considered as wholly different in kind and excluded from ‘standard’ bibliometric analyses.

Collab-CNCI recognizes the different outcomes of multiple authorship and benchmarks the citation count for each paper against a relevant group in the same way as conventional CNCI, whereas fractional counting apportions counts by author numbers and addresses and thereby reduces the value of authorship as collaboration rises. For example, papers that have a trilateral international authorship are judged as worth only one-third of their natural value to an author-country under fractional counting, but under Collab-CNCI are compared with the world average for similarly trilateral papers in the same Web of Science subject category and publication year. The citation value of a paper is not judged and then reduced by fractional apportionment but instead scaled as a relatively good - or poor - paper of its type. This retains the original logic of normalizing citation counts.

Using a defined group of institutions spread globally, we compare Collab-CNCI results with full (conventional CNCI) and fractional counting methods to investigate improvement over a ten-year period. From this, we examine the disaggregated CNCI values of the different components that make up Collab-CNCI. Finally, we review the management value of using comparative CNCI indices and the enhanced information presented by deconstructed profiles.

Methods

The data source was all documents with type article indexed in Web of Science Core Collection (i.e., including the Emerging Sources Citation index ESCI) over a ten-year period from 2009 to 2018. This is the same source used in Potter et al. (2020).

To cover a range of regions and institutional size, as well as minimizing computation time, five institutions were chosen from each of six geo-political regions (North America, Western Europe, Eastern Europe, Middle East-North Africa-Turkey [MENAT], South East Asia, and China) for analysis. Institutions, based on the Enhanced Organization name labels assigned in address data within the Web of Science, were chosen from those who (co)authored at least 5,000 articles over the ten-year period. These chosen institutions produced a dataset covering ~1.2M articles.

The metrics to investigate performance were CNCI, Collab-CNCI, and Fractional CNCI. CNCI was calculated by normalizing an article’s citation count by document type, year of publication, and Web of Science subject category. For an article in multiple Web of Science categories, the average CNCI of these categories was assigned to the article. Each unique institution listed in the article address was allocated full credit. Thus, if an article with two institutions had a CNCI of 1.3, each institution was awarded an article count of 1.0 and a CNCI value of 1.3. An institution’s mean CNCI was then calculated by summing the CNCI of all the articles on which the institution was recorded in an author address and dividing this by the number of such articles, over the entire dataset.

Collab-CNCI, defined by Potter et al. (2020), uses collaboration type as an additional normalization factor. The collaboration types, based on author address data, are divided into domestic: single (authors from the same institution) and multi (authors from more than one institution within the same country), and international: bilateral (two countries), trilateral (three countries), and quadrilateral-plus (four or more countries), regardless of the number of institutions. Each institution on each article was assigned the appropriate collaboration type. As with conventional CNCI, an institution was given full credit and the full CNCI value for an article, though the mean CNCI was calculated considering each collaboration-type separately (e.g., only an institution’s international bilateral articles were considered when calculating the institution’s CNCI for international bilateral).

Fractional CNCI was calculated at the author level based on the method of Waltman and van Eck (2015); this method assigns credit to each author, institution, country etc. based upon the total, deduplicated number of each on an article. The reader is referred to Waltman and van Eck (2015) for additional information.

As this study uses Web of Science data, it should be noted that an equivalent analysis using other indexing services may yield slightly different results.

Results

The calculated CNCI indicators for the chosen ten-year period are listed in Table 1 for each of the five institutions selected in each region along with the total number of articles considered. Since we are interested in trends relating to perceived improvement in citation impact, we provide the CNCI indicator for the first and last year in the study period (2009 and 2018) along with the percentage improvement (%+ column).

Table 1 - Institutional CNCI indicators (CNCI, Collab-CNCI and Fractional CNCI) and their relative improvement (%+ column) over the 10-year period 2009-18.

			CNCI			Collab CNCI			Frac CNCI		
	Institution	**Count**	**2009**	**2018**	**% +**	**2009**	**2018**	**% +**	**2009**	**2018**	**% +**
China	BEIJING NORMAL UNIV	21,503	0.94	1.21	28.8	0.88	1.17	34.0	0.78	1.09	40.3
	HUAZHONG UNIV SCI & TECHNOL	46,547	0.80	1.34	67.5	0.80	1.34	68.1	0.70	1.24	78.8
	SHANGHAI JIAO TONG UNIV	75,800	0.99	1.18	19.2	0.94	1.11	17.9	0.83	0.97	17.3
	TSING HUA UNIV	69,329	1.06	1.45	37.5	1.03	1.33	29.2	0.97	1.22	25.8
	WUHAN UNIV	35,159	0.97	1.32	36.1	0.97	1.26	30.3	0.91	1.16	27.3
MENAT	CAIRO UNIV, EGYPT	19,932	0.74	1.03	38.4	0.71	0.75	6.9	0.75	0.61	-19.0
	ISLAMIC AZAD UNIV, IRAN	60,177	0.64	0.96	49.9	0.64	0.88	37.1	0.62	0.77	22.9
	ISTANBUL UNIV, TURKEY	19,146	0.59	0.75	27.3	0.52	0.57	8.9	0.46	0.44	-4.2
	KING SAUD UNIV, SAUDI ARABIA	29,656	0.73	1.26	73.0	0.64	0.90	41.6	0.58	0.94	63.4
	UNIV TUNIS EL MANAR, TUNISIA	10,974	0.55	0.62	13.5	0.48	0.48	0.6	0.48	0.42	-13.7
E Europe	CHARLES UNIV PRAGUE, CZECHIA	33,534	0.86	1.30	50.4	0.67	0.86	28.0	0.62	0.69	10.7
	EOTVOS LORAND UNIV, HUNGARY	9,148	0.96	1.71	78.0	0.77	1.17	50.8	0.69	0.87	25.6
	JAGIELLONIAN UNIV, POLAND	21,771	0.85	1.36	58.7	0.68	0.91	34.7	0.66	0.79	19.5
	UNIV LJUBLJANA, SLOVENIA	23,060	0.95	1.22	27.6	0.76	0.88	16.1	0.66	0.71	8.5
	UNIV TARTU, ESTONIA	10,168	1.05	1.86	77.4	0.89	1.11	24.2	0.85	0.98	14.5
W Europe	FREE UNIV BERLIN, GERMANY	47,015	1.30	1.66	28.0	1.05	1.17	11.5	1.09	1.14	5.0
	UNIV ANTWERP, BELGIUM	20,172	1.18	1.77	49.6	0.97	1.17	20.2	1.02	1.14	12.3
	UNIV BOLOGNA, ITALY	39,558	1.12	1.67	48.5	0.97	1.14	17.8	0.94	1.00	5.8
	UNIV COPENHAGEN, DENMARK	65,169	1.52	1.83	19.8	1.19	1.26	5.6	1.24	1.29	4.4
	UNIV OXFORD, UK	86,532	1.87	1.87	0.2	1.47	1.36	-7.7	1.58	1.44	-8.9
N America	HARVARD UNIV, USA	192,485	2.13	2.17	1.7	1.87	1.75	-6.4	2.05	1.77	-13.9
	MIT, USA	60,656	2.48	2.24	-9.7	2.28	1.90	-16.7	2.28	1.80	-21.4
	UNIV CALIF SAN DIEGO, USA	61,881	1.87	1.93	3.0	1.66	1.57	-5.2	1.65	1.45	-12.1
	UNIV GEORGIA, USA	27,178	1.22	1.17	-4.4	1.12	1.09	-2.1	1.11	1.06	-4.2
	UNIV TORONTO, CANADA	108,047	1.55	1.81	16.2	1.27	1.26	-0.5	1.33	1.20	-9.6
SE Asia	IISC BANGALORE, INDIA	16,929	0.93	1.02	9.5	0.92	0.91	-1.4	0.87	0.80	-7.1
	MAHIDOL UNIV, THAILAND	14,710	1.03	1.08	4.9	0.81	0.78	-4.3	0.77	0.62	-19.8
	NATL UNIV SINGAPORE	57,666	1.35	1.55	15.5	1.22	1.22	-0.2	1.29	1.31	0.9
	UNIV DELHI, INDIA	12,320	0.76	1.08	42.2	0.72	0.87	21.3	0.70	0.66	-6.6
	UNIV MALAYA, MALAYSIA	27,073	0.62	1.25	102.2	0.60	0.84	41.0	0.53	0.75	40.7

There is significant variation in trends over the period, as would be expected from choosing such a diverse set of institutions. When considering conventional CNCI, institutions in most regions have increased their performance, especially in China, MENAT, Eastern Europe, and South East (SE) Asia where CNCI has often risen from below 1.0 (world average). In Western Europe and North America, some institutions show an increase, some are relatively stable, and others decline slightly.

When reading across to compare Collab-CNCI or Fractional CNCI, a different picture emerges. Institutions from China and Eastern Europe still show apparent increases (albeit smaller), but improvement is suppressed (or in fact, negated) in Western Europe and North America. In MENAT and SE Asia, some institutions have contrary trends when comparing CNCI to Collab-CNCI or Fractional CNCI. For example, the percentage improvement for Cairo University with CNCI is 38.4%, but only 6.9% with Collab-CNCI and -19% with Fractional CNCI. Others, such as Islamic Azad University and University of Malaya, show improvement across all three CNCI indicators.

To illustrate these trends over time, we provide a time-series for the CNCI indicator in Figure 1 (left) for a sample institution in each region. Alongside this (right), the relative growth in internationally co-authored papers is plotted. In all cases, significant growth is seen in the abundance of these collaborative papers, but it is not possible to ascertain from this information whether these papers are responsible for the increase in CNCI.

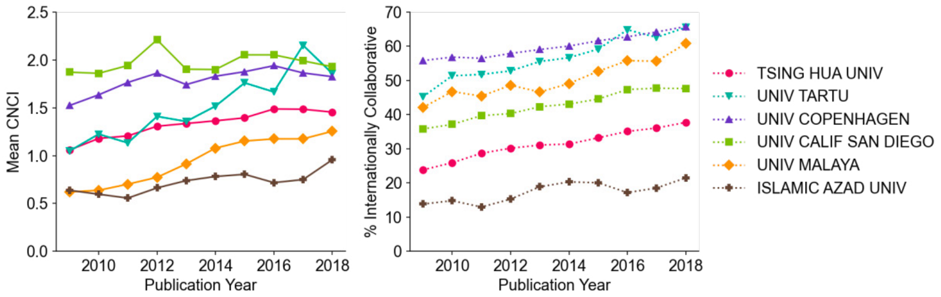


Figure 1 - Mean CNCI (left) and % papers with an international co-author (right) for six sample institutions.

For comparison, a timeseries is plotted in Figure 2 showing the Collab-CNCI indicator (left) and Fractional CNCI (right) over the same time period. Such a plot draws attention to the difference and similarities in trends. For example, Tsing Hua University sees consistent improvement across all CNCI measures, but University of California San Diego changes from relatively consistent performance for CNCI, to a slight decline with Collab-CNCI, and more moderate decline with Fractional CNCI.

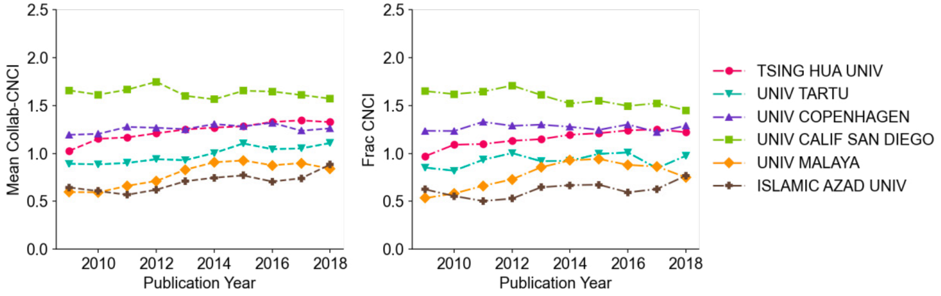


Figure 2 - Mean Collab-CNCI (left) and Fractional CNCI (right) for six sample institutions.

The effect of individual collaboration groups on CNCI values for three sample institutions (Islamic Azad University, Iran; Tsing Hua University, China; University of Copenhagen, Denmark) over the ten-year period is highlighted in Figure 3. As research becomes more collaborative the mean CNCI value increases. This is particularly notable for quadrilateral-plus articles where mean CNCI values are ~1-3 times greater than the other collaborative groups. The quadrilateral-plus articles also have the most variance over the ten years.

The collaboration normalizing effect of Collab-CNCI is clear – the mean for international quadrilateral-plus articles is dramatically suppressed (other international groups are also suppressed) resulting in similar values to the other collaboration types centered approximately around world average (1.0).

The figure also highlights changes in article output between the institutions. The University of Copenhagen, Denmark (the longest established research economy of the three) shows little variation in CNCI over the ten-year period. However, the number and percentage share of quadrilateral-plus articles has doubled (2009 vs 2018) along with a halving of the relative share of domestic single articles. As of 2018, University of Copenhagen had a relatively even share of articles across all groups. Despite its relatively large share of international quadrilateral-plus papers (~23%), this group's (Collab-) CNCI is no better off than its comparators who have a far lower share (<<10%).

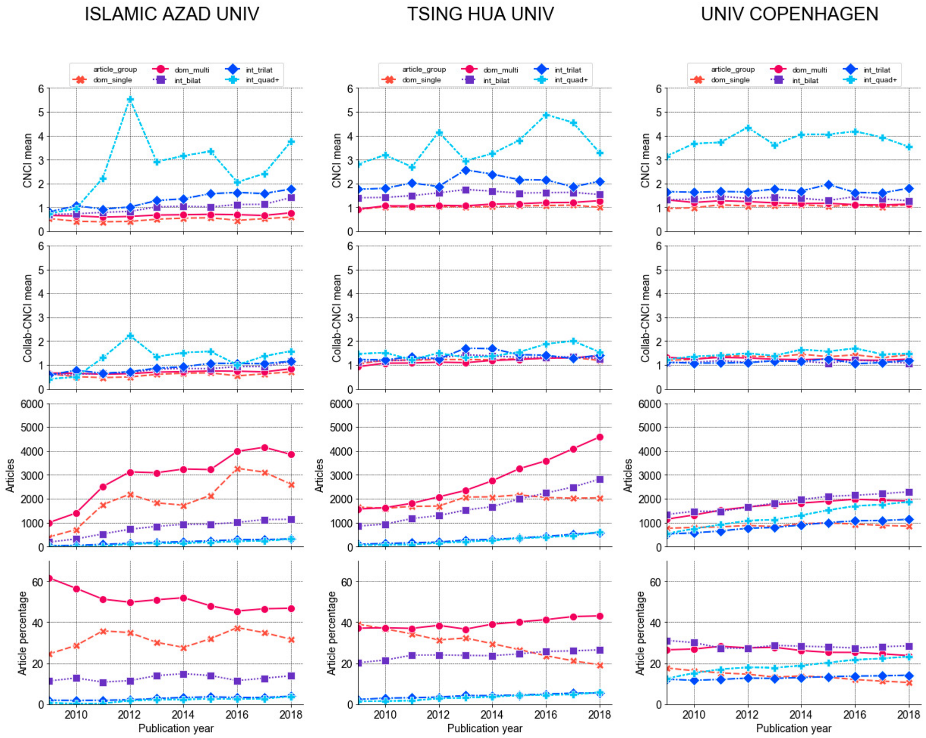


Figure 3 - CNCI and Collab-CNCI mean, and article number and percentage share by publication year for three sample institutions.

Tsing Hua University (China – a growing research economy) shows a two-to-three-fold increase in absolute article number over all collaborative groups, bar domestic single which has remained constant resulting in a relative decrease in its article share (~40% in 2009 vs ~20% in 2018). Domestic multi articles is now the dominant type (~40%), with only incremental changes in international collaboration.

Islamic Azad University (Iran - a developing research economy) has also seen significant increase in article output over the ten years, though in a less consistent manner than Tsing Hua

University. Unlike its comparators, Islamic Azad's domestic single output increased over the period. Domestic multi decreased by a third; international groups again only incrementally improved their share.

Despite their much lower share of international trilateral and quadrilateral-plus articles, Islamic Azad and Tsing Hua have comparable, if not greater, (Collab-) CNCI values for these groups relative to University of Copenhagen. Overall, Copenhagen was the worst improver of these three in terms of Collab-CNCI (5.6% vs 29.2% [Tsing Hua] vs 37.1% [Islamic Azad]).

Discussion

CNCI, while providing a normalized value for institutional comparisons, does not reveal the intricacies of an individual institution's profile, particularly that of collaboration. The Collab-CNCI method does take collaboration into account and can change institutions' metric values (Table 1). Institutions ranked highly by the CNCI metric are more likely to see negative effects in their CNCI improvement when calculated by Collab-CNCI over the ten-year period (e.g., Harvard, MIT, Oxford). This is because such institutions are likely to have strong relative performance across all collaborative groups and already be more internationally collaborative, compared to institutions from developing research economies. This is further highlighted by institutions in Malaysia, Saudi Arabia, Hungary, and China having some of the greatest improvement in Collab-CNCI over the ten-year period. Though well-regarded institutions may show low or negative improvement, when comparing 2009 to 2018, they remain the most highly ranked institutions by Collab-CNCI. Additionally, the (Collab) CNCI improvement metric only considers the raw change in CNCI over the period (i.e., citation pattern changes); interpretation of this in respect to research quality is a separate issue and should be considered along with other performance indicators.

Fractional CNCI, for the chosen sample institutions, provides comparable trends and outcomes to Collab-CNCI (Figure 2). Crucially, however, Collab-CNCI does not assign potentially incorrect credit to each party, while still accounting for collaboration. Due to this, we advocate the use of Collab-CNCI over Fractional CNCI for analysis purposes.

The results further demonstrate the increase in international collaboration over time. In particular, the growth of output in the developing research economies compared to those established ones (Figure 3). Of the individual collaborative groups, international quadrilateral-plus shows the biggest difference between CNCI and Collab-CNCI; the additional normalization demonstrating that these multi-lateral, highly cited articles skew the regular CNCI values. This group also contains the highest variance temporally, due to this group being the smallest by article count (e.g., Figure 3).

The disaggregation of Collab-CNCI into the constituent groups provides an additional level of information and insight that cannot be gained from the single metric value. This, therefore, allows research managers and funders to better contextualize relative performance across and within institutions and, hence, make better informed funding decisions.

Acknowledgments

The authors are grateful to colleagues at the Institute for Scientific Information and Clarivate for their advice and support in developing this analysis.

References

Adams, J. (2012). Collaborations: the rise of research networks. *Nature*, 490, 335-336.

Adams, J. (2013). The fourth age of research. *Nature*, 497, 557-560.

Adams, J. & Gurney, K. A. (2018). Bilateral and multilateral coauthorship and citation impact: patterns in UK and US international collaboration. *Frontiers in Research Metrics and Analytics*, 23 March. https://www.frontiersin.org/articles/10.3389/frma.2018.00012/full

Adams, J., Pendlebury, D.A., Potter, R.W.K., & Szomszor, M. (2019). Multi-authorship and research analytics, Clarivate Analytics, London UK. ISBN 978-1-9160868-6-9

Aksnes, D.W. (2003). Characteristics of highly cited papers, *Research Evaluation*, 12(3), 159–170, https://doi.org/10.3152/147154403781776645

Aksnes, D.W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: an overview of basic concepts and theories. *Sage Open*, 9 (1), 1-17.

Bozeman, B., Fay, D., & Slade, C.P. (2013) Research collaboration in universities and academic entrepreneurship: the state-of-the-art, *Journal of Technology Transfer*, 38, 1-67.

Evidence (2007). The use of bibliometrics to measure research quality in UK higher education institutions. Report to Universities UK. Universities UK, London. ISBN 978 1 84036 165 4 https://dera.ioe.ac.uk//26316/ (accessed 17 Sept 2020).

Garfield, E. (1955). Citation indexes for science. A new dimension in documentation through association of ideas. *Science*, 122, 108-111.

Jappe, A. (2020). Professional standards in bibliometric research evaluation? A meta-evaluation of European assessment practice 2005–2019. *PLoS One*, published: April 20, 2020. https://doi.org/10.1371/journal.pone.0231735

Katz, J.S. & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, 40(3), 541-554

Leydesdorff, L. & Wagner, C.S. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics*, 2(4), 317-325.

Potter, R.W.K., Szomszor, M., & Adams, J. (2020). Interpreting CNCIs on a country-scale: The effect of domestic and international collaboration type. *Journal of Informetrics*, 14(4), 101075.

Szomszor, M., Adams, J., Fry, R., Gebert, C., Pendlebury, A.D., Potter, R.W.K., & Rogers, G. (2021) Interpreting bibliometric data. *Frontiers in Research Metrics and Analytics*, 09 February 2021. doi: 10.3389/frma.2020.628703

Wagner, C. S. (2008). The New Invisible College. Washington, DC: Brookings Press.

Wagner, C.S. & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34, 1608-1618.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, 10, 365–391. doi: 10.1016/j.joi.2016.02.007

Waltman, L. & van Eck, N.J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics*, 9(4), 872-894. doi 10.1016/j.joi.2015.08.001

Waltman, L. & van Eck, N.J. (2016). The need for contextualized scientometric analysis: An opinion paper. In Ràfols, I., Molas-Gallart, J., Castro-Martínez, E., Woolley, R., (Eds)., Proceedings of the 21st international conference on science and technology indicators, València: Universitat Politècnica de València, pp. 541–549.

Five decades of empirical research on the postdoc: a scoping review of the contribution of bibliometrics

Heidi Prozesky[1]

[1] *hep@sun.ac.za*
DST-NRF Centre of Excellence in Scientometrics and STI Policy; and Centre for Research on Science and Technology (CREST), Stellenbosch University, Ryneveld Street, 7600, Stellenbosch (South Africa)

Abstract

Although variety of concerns with the postdoc have been growing internationally, scoping or systematic reviews of research on the topic have been limited to a single field and/or a particular concern, and none have considered publication performance per se, which is arguably a central concern of the postdoc. I address this gap through a scoping review of empirical evidence on the postdoc in general, as published over the past five decades, in terms of volume and research design. I also determine, both quantitatively and qualitatively, the contribution of bibliometrics – as an approach to measurement of publication performance – to that body of evidence. All document types indexed in Web of Science and/or Scopus (1970–2020) were searched for spelling variants/derivatives of "postdoc" in the title field. Manual inspection of search results and full-text versions applied predetermined definitions of "the postdoc" and "empirical research". Data on the resulting 260 sources of evidence were analysed, showing an increase in the volume, and the dominance of surveys and secondary analysis, which renders systematic reviews of these feasible. This does not apply to bibliometrics, which has made a small contribution, mostly as part of evaluation studies, and subject to methodological limitations and knowledge gaps.

Introduction

Although there is a lack of a systematic, global definition of the postdoctoral appointment, or "postdoc" (Åkerlind, 2005), for the purposes of this paper it is defined as a temporary position, after completion of a doctorate, taken primarily for additional training, in particular advanced research apprenticeship under supervision (Gaughan & Bozeman, 2019). The earliest postdoctoral appointments (which were in the United States) have been traced back to the 1870s (Zumeta, 1985). In other countries, the postdoc is a comparatively new phenomenon, but viewed globally, the postdoc as an institution has become increasingly common in the past few decades. Its growth has not been without concerns about, for example, their potential or actual status of precarity, and their status as neither staff nor students. There is, however, consensus that postdocs are first are foremost researchers (or at least should be), thus their research production – before, during and after their position – should also be a central concern.

The rationale for this scoping review was informed more generally by my interest in studying the postdoc in South Africa – through a survey and bibliometrically – and specifically by my realisation that only five scoping or systematic reviews have been conducted on the topic internationally (Numminen et al., 2020; Treichler et al., 2020; Nowell et al.; 2019; Nowell et al., 2018; Ranieri et al., 2016). Moreover, these were limited to one field and/or to a particular issue, and did not consider research production per se.

To address these knowledge gaps, I conducted a scoping review of the available empirical evidence on the postdoc, across fields and topics, to determine (1) its volume over the past five decades (1970–2020); and (2) the designs employed to produce the evidence. As a precursor to future systematic reviews, this quantitative part of my scoping review allowed me to assess the feasibility of such future initiatives, but also to determine the extent to which postdoc-related publication performance has been studied using bibliometric data.

I then performed a more in-depth analysis of the instances of bibliometrics this identified, by employing both quantitative and qualitative content analysis to understand, in more depth, why and how bibliometric data were collected and analysed, and in which context (country and field). However, following Peters et al.'s (2020, p. 421) guidelines, I did not undertake thematic analysis or evidence synthesis, as this would be "beyond the scope of a scoping review". Rather, my focus was on methodological limitations as well as possibilities, and knowledge gaps in terms of context studied, that may be addressed by future bibliometricians interested in studying the postdoc.

Methods and procedures

Inclusion criteria

The core concept examined by the scoping review is the "postdoc", and the following definition of a postdoctoral researcher – agreed upon by the United States of America (USA) National Postdoctoral Association, National Institutes of Health, and National Science Foundation – informed my inclusion criteria: an "individual who has received a doctoral degree (or equivalent) and is engaged in a temporary and defined period of mentored advanced training to enhance the professional skills and research independence needed to pursue his or her chosen career path" (Institute of Medicine, 2014, p. 3). I kept in mind that this USA-based definition does not take account of cross-national differences in the definition of the postdoc, as highlighted by Åkerlind (2005), but (as detailed below) it emerged as the most useful operational definition to apply. My review considers not only research on postdocs as individuals, but on issues, other individuals and/or institutions related to "the postdoc".

Based on the assumption that sources of evidence that I was interested in reviewing – those focusing exclusively, or primarily, on the postdoc – would include this term in the title, I searched for two accepted spelling variants, as well as the derivatives, of the term (expressed as the string "postdoc* OR post-doc*") in the (article) title field. As I was interested in empirical evidence of comparable quality, and working within constraints of time and resources, only the Web of Science (WoS) Core Collection and Scopus were searched. These two "world-leading […] citation databases" are "widely used in meta-analysis related studies" (Zhu & Liu, 2020, p. 321). Because of its somewhat broader coverage (Aksnes & Sivertsen, 2019; Mongeon & Paul-Hus, 2016), Scopus was used to augment the results of the WoS search.

However, I recognise that both databases have deficiencies in relation to the social sciences and humanities, and the coverage of literatures in other languages than English. Also, I deviated from Peters et al.'s (2020) three-step search strategy for scoping reviews, namely (1) an initial limited search of at least two appropriate online databases relevant to the topic; (2) a second search, using keywords and index terms identified from the first search, across all included databases; and (3) a search of the reference list of identified sources, for additional sources.

To counteract the potential limitations of my search strategy, I kept my review as "open" (Peters et al., 2020, p. 410) as possible, by placing no limitation on document type, language, contextual setting (e.g., discipline, country) or other criteria. Rather, I applied (manually, to full-text versions) a definition of "empirical research" as conforming to a standard logic, namely the ProDEC framework. "ProDEC" refers to the four elements that are standard in all forms of empirical research: a research problem (Pro), research design (D), empirical evidence (E) and conclusions (C) (Babbie & Mouton, 2001). If at least three of these elements were reported by authors, who themselves conducted the research (i.e., excluding science journalists' reports, usually in trade journals), the source was included.

Sources of evidence published before 1970 were excluded, because WoS searches are limited to publications published from 1970 onwards, and Scopus includes references on publications back to the same year. In addition, postdoctoral training had not been fully institutionalised prior to 1970 (Reskin, 1976). Sources of evidence published up to the end of December 2020 (including early-access publications) were searched, and the time frame is therefore exactly 51 years (approximately five decades).

Search strategy, source selection, full-text retrieval and data extraction

Starting in March 2020, WoS was searched first. Each of the 1 296 results of the search strategy were examined, and two main types of sources were excluded: (1) "noise", e.g., "postdocetaxel"/"post-docetaxel" in the title; and (2) sources that could clearly and unequivocally be considered non-empirical, based on a relatively brief examination (mainly calls for applications for postdoctoral fellowships; announcements of successful applicants for such fellowships; and book reviews). For the remainder of the 444 sources, the following two actions were performed: (1) the full-text was retrieved or requested through an interlibrary-loans service (hard copies received were scanned); and (2) relevant data on each source were extracted from WoS and entered in an Excel spreadsheet. The process was repeated for Scopus, which produced 1 288 results, the majority of which were duplicates of the WoS results. After exclusion of these and the two types on sources referred to above, full-text retrieval and data extraction were undertaken for the remaining 367 sources. These processes were repeated in January 2021, to ensure inclusion of as many sources published in 2020.

The next stage involved assessing the 811 sources in the data set to determine whether they met above-mentioned, predetermined inclusion criterion of "empirical research". The classification was based on the extraction of methodological and other data from the full-text versions that had been successfully retrieved for all but three of those sources (two Portuguese articles and a German one). For these three sources, as much methodological data as possible were extracted from their English abstracts. During full-text examination, the data set was further refined, through the exclusion of 31 sources because they either (1) employed a wider definition of "postdoc/toral" than the one I chose to apply (e.g., academics with a PhD; the general career stage after attaining a PhD); or (2) reported on research concerning postdocs and other groups (e.g., postgraduate students; junior faculty), and did not disaggregate the postdocs from these other groups.

Of the remaining 780, I classified 260 (33%) as sources of evidence produced by empirical research. The majority (220, or 85%) are articles, and the remaining 40 are comprised of relatively similar numbers of editorial material (12); conference papers (9); reviews (8); and book chapters (7). Two books (or, rather, committee reports), one meeting abstract (comprising 13 pages) and one "note" (comprising 26 pages) were also included. The majority of the items (210, or 81%) were sourced from WoS. Six are not written in the English language, but sufficient methodological data could be collected for accurately coding of research design of all but one of these, which was excluded from the analysis of design.

This explicit (and rather lengthy) description ensures that my review is transparent and reproducible. However, it should be noted that I was the only reviewer, while Peters et al. (2020) require that source selection should be performed by two or more reviewers, independently.

Data processing and analysis

One of the aims of my quantitative analysis was to identify and determine the proportion of sources of evidence on the postdoc that were produced, at least in part, by bibliometric research. I therefore classified designs inductively, through immersion with the full-text versions, focusing mainly on the type of data collected and/or analysed. Designs such as case studies, action research, field experiments, mixed methods designs and evaluation studies tend to collect and/or analyse data from various sources, and were therefore coded according to the combinations of types of data collected and/or analysed. In the case of evaluation studies, I focused on the methods used to measure outcomes of an intervention. It should also be noted that this design category (re-)emerged as useful in the in-depth analysis of the sub-set of bibliometric sources of evidence.

Approximately a fifth of the sources report on a wide variety of combination of two or more designs, therefore the individual designs, and not the sources of evidence, were taken as the unit of analysis. After coding had been completed, the data were exported to IBM SPSS Statistics (version 27), to conduct quantitative, univariate analysis. Based on this analysis, the full-text versions of the instances of bibliometric research could be identified, and these were subjected to quantitative and qualitative content analysis, with a focus on when, why, how and in which context (country and field) the bibliometric evidence was produced.

Findings

Quantitative results of an analysis of sources of evidence on the postdoc

The 260 sources of evidence on the postdoc were published over 48 years (1973–2020). Table 1 summarises the distribution of the sources over these (roughly) five decades, showing that two-thirds of the sources of evidence were published in the past 11 years. Percentages decline quite rapidly as one moves back in time, and are very low prior to 1990. On average, in the 1970s and 1980s less than one source was published per annum, increasing to 2,5 in the 1990s, and doubling to approximately five in the 2000s. In the period 2010–2020, however, the average increases more than three-fold, to 15,8 sources published. The past six years were particularly productive in terms of evidence on postdocs published: 2018 leads with 29 publications, followed by 2020 (26), 2016 (23) and 2015 (22).

Table 1. The distribution of sources of evidence (n=260) published in the period 1973–2020.

Year range	*Frequency*	*Percent*	*Mean*
1973-79	6	2%	0,9
1980s	8	3%	0.8
1990s	25	10%	2,5
2000s	47	18%	4,7
2010-20	174	67%	15,8
Total	**260**	**100**	

Figure 1 summarises the 320 designs that were applied in 259 of the sources of evidence in my data set.

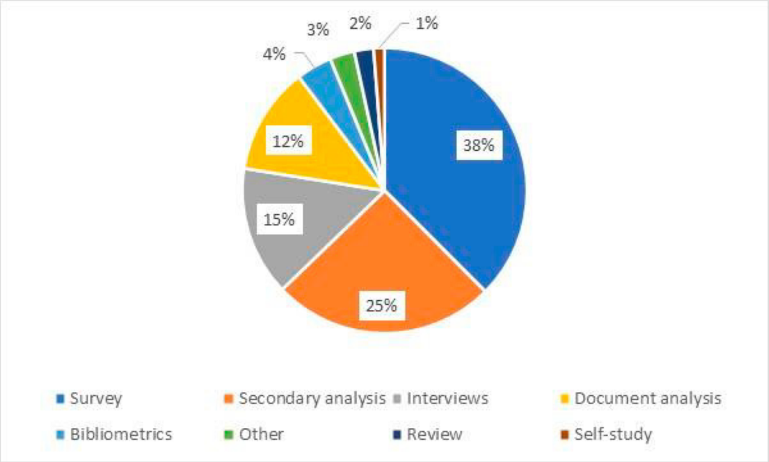


Figure 1. Distribution of research designs (n=320 designs)

As I derived my classification from the data itself, rather than using preconceived categories, some clarification of the categories is provided as part of the results. Of the 320 individual designs that were reported, more than a third (38%) are surveys, using questionnaires (or in very few cases, structured interviewing) to collect quantitative or quantifiable data on more than one case. Another quarter (25%; n=81) of the designs involved the secondary analysis of existing data (or statistics) collected by institutions (ranging from national governments to administrators of a funding programme), or researchers other than the authors. Interviews (n=47) were third-most frequently applied (15%). These ranged from semi-structured to in-depth interviews with a relatively small number of participants, and were coded as such if the data and analysis were primarily qualitative in nature.

Document analysis constitutes 12% (39) of the designs. The term "document" covers a very wide range of different kinds of source, but following Bryman (2012), they can be classified as follows: personal documents [e.g., curriculum vitae (CVs) and funding applications]; official documents deriving from either the state (e.g., policy reports and legislation) or private sources (e.g., funders' guidelines or requirements); virtual outputs (e.g., websites and social-media platforms); and mass-media outputs (e.g., job advertisements).

This defined, document analysis (or even of secondary analysis) may include bibliometrics, but I distinguished it as a unique design, if it conformed to the definitions of Thompson and Walker (2015, p. 551) and of Godin (2006, pp. 109–110), namely "the application of mathematical and statistical methods to scholarly publications" to measure the "output side of science". Only 4% of the designs met this definition, and these 13 instances will be described in more detail in the next sub-section of my findings.

The remaining designs are "review" (a sub-set of document analysis) and "self-study". The seven instances of reviews include the five scoping or systematic reviews on the postdoc that I referred to in my introduction, as well as two (Towns et al, 2017, and Sherry et al., 2103) that do not meet the criteria of such reviews but were used to collect empirical evidence. Self-study (n=4) is "a reflective, analytical approach to understand the authors' individual and shared experiences, practices, and insights" (Ovens & Fletcher, 2014, as cited in Nowell, Grant & Mikita, 2019, p. 306). Nine designs were too unique to be classified into one of these seven categories, and were thus coded as "other". The remainder of this paper provides the results of a content analysis of those 13 sources of evidence on the postdoc that involved, at least in part, a bibliometric component.

Results of a content analysis of bibliometric sources of evidence on the postdoc

Only two of the 13 sources of bibliometric evidence were not published as articles, but as editorial material (specifically a commentary) (Mbuagbaw et al., 2018), or a peer-reviewed conference paper (Zabetta & Geuna, 2019). All but one (De Castro & Porto, 2012) were sourced from WoS, and for this article, only the abstract is available. The summary in Table 2 below shows that, in line with the sources of evidence on postdocs in general, the majority (n=9) of the bibliometric sources were produced in the past decade. After Reskin's 1976 article, almost 30 years passed before the postdoc was again analysed bibliometrically (and then only three times in the 2000s).

The majority (n=10) of the sources may be classified as evaluative bibliometrics (van Leeuwen, 2004), and of these, most (n=7) evaluate one, or a small number of similar, interventions. In these cases, bibliometrics is used to primarily to determine their publication-performance outcomes. These interventions are (in order of earliest to latest publication date): the Burroughs Wellcome Fund Career Awards in the Biomedical Sciences (Pion & Ionescu-Pioggia, 2003); the Canadian Association of Gastroenterology – Canadian Institutes of Health Research – pharmaceutical partner postdoctoral operating fellowship programme (McKay & Daniels, 2003); the Boehringer Ingelheim Fonds (Bornmann & Daniel, 2006); the NIH National Institute of General Medical Sciences Kirchstein-NRSA F32 grant (Levitt, 2010) and the more general NIH F32 grant (Jacob & Lefgren, 2011); two funding instruments of the Council for

Independent Research in Denmark, focussing on three research councils (Schneider & van Leeuwen, 2014); the CIHR Canadian HIV Trials Network international postdoctoral fellowship for capacity building in HIV clinical trials (Mbuagbaw et al., 2018); and three prestigious NASA fellowships (Hubble, Einstein and Sagan) (Pepper et al., 2019). The exceptions are Reskin's study of the impact of postdoctoral fellowships on productivity in one field, and two studies on the effect of appointments or training abroad, i.e., outside of Portugal (De Castro & Porto, 2012) and Italy (Zabetta & Geuna, 2019).

Table 2. Summary of features of sources of evidence (n=13) produced by bibliometric analysis (1976–2019)

Author(s)	*Pub. date*	*Time frame*	*Sample size (publications)*	*Country*	*Discipline/field*
Reskin	1976	1955–1970	450	USA	Chemistry
Pion & Ionescu-Pioggia	2003	1995–2000	101	USA & Canada	Biomedical sciences
McKay & Daniels	2003	1992–2000	87 (247)	Canada	Gastroentorology & related disciplines
Bornmann & Daniel	2006	1990–1995	397 (1 586)	Germany	Biomedical sciences
Levitt	2010	1992–2009	439	USA	Basic life sciences
Jacob & Lefgren	2011	1980–2000	12 189 (13 426)	USA	Biological sciences (predominantly); physical, social & "miscellaneous"
De Castro & Porto	2012	Unknown	86	Portugal	Biology, engineering, geosciences & health sciences
Miller & Feldman	2014	1993–2006	N/A	USA	Life sciences
Schneider & van Leeuwen	2014	2001–2009	632 (6 196)	Denmark	Health, natural & technical sciences
Igami, Nagaoka & Walsh	2015	2001–2006	4 410 (≈10 000)	USA & Japan	Natural sciences
Mbuagbaw et al.	2018	2010–2015	6 (40)	Canada	HIV (clinical)
Pepper et al.	2019	2014–2017	133	USA	Astronomy
Zabetta & Geuna	2019	1986–2015	15 385 (285 283)	Italy	All

The interventions that were studied differ widely in scope (i.e., number of applicants and/or awardees), and all 13 sources of evidence differ widely in terms of the time frame studied. The bibliometric data that were analysed range over six decades (1955–2017); on average (and excluding one unknown time frame), the studies cover 11 years, with a maximum of 29 (Zabetta & Geuna, 2019), and a minimum of three (Pepper et al. 2019). Thus, it is not surprising to find great variability in terms of sample size. Excluding one study that analysed universities (Miller & Feldman, 2014), the sample sizes range from a minimum of 40 publications produced by six postdoc recipients of a single fellowship (Mbuagbaw et al., 2018), to a maximum of 285 283 publications of 15 385 individuals (Zabetta & Geuna, 2019). The median is ≈400. Notably, three of the studies (Pepper et al., 2019; Bornmann & Daniel, 2006; Levitt, 2010) focused on an "elite" population, and one on fast-moving and competitive scientific research (Igami, Nagaoka & Walsh, 2015).

In evaluations, the impact of these interventions on subsequent scientific productivity is a central, bibliometrically evaluated concern. However, the methodological rigour (especially a comparison or control group) required to determine such impact is lacking in at least three evaluations. McKay and Daniels (2003) conclude, based on very little evidence, that the fellows of the programme they evaluated "have performed admirably" in terms of both publication rate and publication quality, and – as the title of their article indicates – they assessed the programme as "an outstanding success that continues to excel!" (p. 437). Similarly, in Mbuagbaw et al.'s (2018) "commentary" (editorial material), the 23 authors conclude, also on the basis of scant bibliometric evidence, that the fellows have built a strong network of collaboration and scientific productivity. Pion and Ionescu-Pioggia (2003) is the third instance of such a practice, but they recognise the limitations of lacking an appropriate comparison group, and temper the veracity of their conclusions accordingly.

Three years later, Bornmann and Daniel (2006) addressed this issue by comparing not only successful and unsuccessful applicants, but also their population's values with international scientific reference values. Jacob and Lefgren (2011) followed suit, with their inclusion of both successful and unsuccessful applicants. Importantly, they warn against "naïve comparisons" in this regard, as these "may be biased upward, reflecting both the causal impact of receiving a fellowship as well as differences in latent scientific productivity (or interest)" (p. 866). Five years later, Schneider and van Leeuwen (2014) again improve in terms of this and other methodological issues involved in the evaluation of postdoc-related interventions. Their conclusion is different from the official conclusion given in an evaluation report, which emphasises the success of the funding programmes (reminiscent of McKay and Daniels, 2003, as well as Mbuagbaw et al., 2018). All three of these studies provide valuable descriptions of empirical strategies that have been used by the authors to increase the robustness of findings.

A need for a "logic of comparison" (Bryman, 2012, p. 58) was addressed somewhat differently by De Castro and Porto (2012), who compared scientific production before and after postdoc training, taking account journal quality, and whether the training was abroad or local. Similarly, Zabetta and Geuna (2019) compared internationally mobile postdocs with a matched control group of their non-internationally mobile counterparts. Interestingly, three studies also compare women and men (Reskin, 1976; Levitt, 2010; Pepper et al., 2019). In addition to the evaluation studies, two econometric analyses utilised bibliometric data to study the contribution of postdocs to research productivity across institutions or countries. The first (Miller & Feldman, 2014) considered research productivity in more general terms, while Igami, Nagaoka and Walsh (2015) focused on postdocs contribution to fast-moving and competitive scientific research.

Next, I consider sources of bibliometric data analysed by the 11 publications for which these were specified. The sources are relatively standardised, with at least half using WoS (or its previous designations, such as SCI and ISI Web of Knowledge, including Journal Citation Reports). Other sources include Chemical Abstracts (1976), PubMed (Levitt, 2010), Google Scholar (Mbuagbaw et al., 2018), Scopus (Zabetta & Geuna, 2019) and the NASA Astrophysical Data System (Pepper et al., 2019).

Two other, less-common sources of bibliometric data deserve mention. Miller and Feldman (2014) sourced average citations per publication from the USA's National Research Council Assessments of research doctoral programmes, to weight publication counts. Pion and Ionescu-Pioggia (2003) collected citation data from SCI, but considered the resulting citation rates difficult to interpret, because "no clear standard exists for judging young faculty's performances on this measure" (p. 184). They therefore also measured the percentage of publications in top-ranked journals, not identified as such bibliometrically, but "by intramural NIH scientists as the 23 most prestigious journals in which to publish basic biomedical and clinical research papers" (Hooper, 1984, as cited in Pion & Ionescu-Pioggia, 2003, p. 184).

Importantly, the bibliometric data that were collected were not only used to measure scientific productivity. A prime example is Zabetta and Geuna (2019), who used (1) affiliation data to deduce a postdoctoral appointment abroad and to build a variable on home country linkages; and (2) publications and citations during the PhD as two of six characteristics on which they matched a treatment group with a control group. Igami, Nagaoka and Walsh (2015) used bibliometric data to measure speed of advancement of research, in terms of citation time lag.

If bibliometric data were used to measure scientific productivity, the unit of analysis was not always the postdoc, as illustrated by Levitt's (2010) study of the influence of mentors' scientific on postdocs careers. However, some uses of bibliometric data involve some precarious assumptions. Zabetta and Geuna (2019) used only affiliation data to identify "researchers who, after the PhD and before the first appointment in Italian academia" published with a non-Italian affiliation for "some time", (p. 4), and equated these academics with those that "undertook a postdoctoral appointment before entering the Italian academic system" (p. 2). Similarly, Levitt (2010) assumed a postdoc's mentor to be the senior author on the publications of the postdoc from the fellowship institution during or shortly after the period of the fellowship.

All the studies depended on other data sources for sample identification and/or further analysis. Some of these data were publicly available, for example online websites and CVs (Pepper et al., 2019), and NIH administrative records (Levitt, 2010; Jacob & Lefgren, 2011). Pepper et al. (2019) used CVs also to verify publication records – an interesting variation on the practice of contacting researchers to validate publication lists, which is considered normal in the design of a bibliometric analysis (van Leeuwen 2007, as cited in Schneider and van Leeuwen, 2014, p. 292). Lists of publications were also verified against online databases (Bornmann & Daniel, 2006). In two other cases, the researchers used data they themselves had collected through previous surveys (Igami, Nagaoka & Walsh, 2015; Zabetta & Geuna, 2019), to create an initial sample.

In a third set of sources of evidence, the research was only possible because of access to data that are not publicly available. One case in point Schneider and van Leeuwen (2014), whose bibliometric analysis formed part of a study they were commissioned, by the Council for Independent Research in Denmark, to conduct. Thus, the authors had access to all postdoc grants from 2001 to 2009, including grant information, names, and demographic data (compiled by the Danish Ministry of Science and Innovation), from which they drew their sample of (matched) funded postdocs. They further mention that they had access to data on a general population of researchers in Denmark for the same period, from which they sampled their control group. In other cases, the sources of non-bibliometric data are vague, and it may be deduced that their availability was dependent on the authors' positionality as evaluators of interventions. For example, CVs that are submitted annually by individuals with active awards from a programme were available to an author who is senior programme officer (Pion & Ionescu-Pioggia, 2003).

Finally, I consider the context (country and field) on which the evidence was produced. As international mobility is often associated with the postdoc, it is sometimes difficult to code studies on the postdoc according to this variable. Other information such as the non-bibliometric data sources used, the country/ies a programme supports, and citizenship or permanent residency criteria that apply to applicants were taken into account. The results show that the bibliometric analyses focused most often on the USA, European countries (Germany, Portugal, Denmark and Italy), and Canada. However, in the case of one "Canadian" study (Mbuagbaw et al., 2018), the six fellows studied are mainly located in African countries (Cameroon, Lesotho, South Africa, Uganda and Zambia), as well as in China. In terms of field, the medical and life sciences dominate, followed by the natural sciences. Only two studies (Jacob & Lefgren, 2011; Zabetta & Geuna, 2019) included other fields, such as the social sciences.

Discussion

This paper reports on the results of the first scoping review of (1) sources of empirical evidence on postdocs across fields and topics; and (2) the contribution of bibliometrics to that evidence. The sources in my data set as a whole range from 1973 to 2020, increasing in small increments from a very low base in the first two decades to the next two, but followed by a significant increase in volume over the past decade. This growth pattern differs from the "exponential growth of global scientific publication output" as measured for the period 1980 to 2012 (Bornmann & Mutz, 2015:2217), and its relationship with wider other developments in science systems internationally would be an interesting avenue for future research.

The numbers of sources of evidence on the postdoc seem to indicate feasibility for systematic review. However, if a particular topic or field were to be selected, the numbers would be quite low, as illustrated by previous reviews. Of Numminen et al.'s (2020) 44 studies of competencies required for a postdoctoral nursing researcher career, the majority concerned not postdocs, but PhD students; Treichler et al. (2020) identified only one study on diversity and social justice training of psychology postdocs, and had to broaden their search to include literature on graduate students in other, closely related fields; Nowell et al. (2019, 2018) found slightly less than 30 articles on professional learning and development initiatives for postdocs; and although Ranieri et al.'s (2016) search resulted in 50 articles on factors that influence career progression among postdoctoral clinical academics, their definition of postdocs is quite broad. Importantly, none of these studies used research design as part of their inclusion criteria.

Were one to do so, without focusing on a particular field and/or topic, the number of surveys on the postdoc are probably adequate for a systematic synthesis of the evidence they produced, and this number may be increased by including some of the secondary analyses of existing survey data. The extent to which this is the case, and the comparability of the survey data in terms of population surveyed, require further analysis that is beyond the scope of this paper.

With regard to the other designs, including bibliometrics, systematic reviews of the evidence produced by these designs would not be feasible at this stage. My search was limited to two databases, did not involve searching the reference lists of publications identified, and additional search terms (e.g. early career scientists) would have probably increased the volume of evidence. However, it is unlikely that a more comprehensive search would increase the volume of bibliometrically produced evidence to a level that allows for systematic evidence synthesis. Such a synthesis would also be hampered by wide variations in scope and the fact that context-specific evaluations of postdoc-related interventions dominate the body of evidence reviewed. Among these instances of evaluative bibliometrics, some methodological standardisation (and improvements) has emerged over time, but it is of concern that bibliometric analyses were still used recently to make unfounded conclusions about the "effect" or "impact" of interventions on postdocs' publication performance. Schneider and van Leeuwen (2014, p. 296) even question the contribution bibliometrics can make to such evaluations, on the basis that "publication activity is a poor parameter for performance evaluation in this context". However, innovative uses of bibliometric data to measure more than publication performance, as illustrated by Zabetta and Geuna (2019) and Igami, Nagaoka and Walsh (2015), shows that bibliometric data and tools can be applied to more than the performance evaluation of postdocs. Although bibliometric analyses contributed only a small fraction of empirical knowledge on postdocs over the past five decades, they are not a recent phenomenon, with the earliest already published in 1976 (analysing data as far back as the 1950s). As publication performance is both a major postdoc and bibliometric concern, what has been limiting the volume of such analyses, when compared to surveys and secondary analysis? One possible explanation is the “meagre” publication history of most postdocs around the time they received their grants (within one or two years after finishing their PhD) (Schneider & van Leeuwen (2014, p. 290). As illustrated by some of the sources of evidence I reviewed, studying postdocs over longer time periods, or as part of mentor-mentee dyads or research teams, would allow for larger numbers of publications to be analysed.

Lack of non-bibliometric data may also be a contributing factor. The results of my content analysis show, quite predictably, that all of the bibliometric studies on the postdoc required access to additional data for sample identification, verification of publication data, and analysis. Although some of these data are publicly available, in most cases the authors were privy to these data because of their positionality as evaluators of interventions. However, my scoping review shows that a large majority of empirical evidence on the postdoc since 1973 has been produced by surveys, leading me to question whether availability of sampling frames from which to construct publication portfolios is a valid reason for the low number of bibliometric analyses.

In addition to the dominance of evaluations of specific interventions, the study of elite populations, and of gender differences, were two interesting themes that emerged. Finally, the context to which the bibliometric evidence pertains, tends to reflect the countries and fields in which the postdoc has been institutionalised for the longest time, namely the medical and life sciences, and the natural sciences, in North America and Europe (Assmus, 1993). Postdoc-related bibliometric research on other countries, and on the social sciences and humanities, is a knowledge gap highlighted by my content analysis.

Conclusions

A central concern with the postdoc is publication performance, the objective measurement of which is the sine qua non of bibliometrics. My review shows that bibliometric data can also be used to measure other issues related to the postdoc. However, compared surveys on the postdoc, the body of evidence that has been produced by bibliometrics since 1976 is still insufficient in terms of volume, standardisation and rigour, for systematic evidence synthesis to be undertaken. In efforts to increase the bibliometrically produced evidence-base on postdocs, current biases towards the medical and life sciences, and the natural science, in North America and Europe should be addressed.

Acknowledgments

Funding from the DSI-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy is hereby acknowledged.

References

Åkerlind, G.S. (2005). Postdoctoral researchers: roles, functions and career prospects. *Higher Education Research & Development*, 24, 21–40.

Aksnes, D.W. & Sivertsen, G. 2019. A criteria-based assessment of the coverage of Scopus and Web of Science. *Journal of Data and Information Science*, 4, 1–21.

Assmus, A. (1993). The creation of postdoctoral fellowships and the siting of American scientific research. *Minerva*, 31, 151-183.

Babbie, E. & Mouton, J. (2001). *The Practice of Social Research* (South African Edition). Cape Town: Oxford University Press South Africa.

Bornmann, L. & Daniel, H-D. (2006). Selecting scientific excellence through committee peer review: A citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, 68, 427–440.

Bornmann, L. & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66, 2215–2222.

Bryman, A. (2012). *Social Research Methods* (4th ed.). Oxford: Oxford University Press.

De Castro, P.M.R. & Porto, G.S. (2012). Evaluation of outputs of post-doctoral research: Brief notes about the scientific production in journals. *Ensaio*, 20, 51–72.

Gaughan, M. & Bozeman, B. (2019). Institutionalized inequity in the USA: The case of postdoctoral researchers. *Science and Public Policy*, 46, 358–368.

Godin, B. (2006). On the origins of bibliometrics. *Scientometrics*, 68, 109–133.

Igami, M., Nagaoka, S. & Walsh, J.P. (2015). Contribution of postdoctoral fellows to fast-moving and competitive scientific research. *The Journal of Technology Transfer*, 40, 723–741.

Institute of Medicine (2014). *The Postdoctoral Experience Revisited*. Washington, DC: The National Academies Press. Retrieved 21 February, 2020 from: https://doi.org/10.17226/18982.

Jacob, B.A. & Lefgren, L. (2011). The impact of NIH postdoctoral training grants on scientific productivity. *Research Policy*, 40, 864–874.

Levitt, D.G. (2010). Careers of an elite cohort of U.S. basic life science postdoctoral fellows and the influence of their mentor's citation record. *BMC Medical Education*, 10, 80.

Mbuagbaw, L., Slogrove, A.L., Sas, J., Kunda, J.L., Morfaw, F., Mukonzo, J.K., Cao, W., Ngomba-Kadima, G., Zunza, M., Ongolo-Zogo, P., Nana, P.N., Cockcroft, A., Andersson, N., Sewankambo, N., Cotton, M.F., Li, T., Young, T., Singer, J., Routy, J-P., Ross, C.J.D. Thin, K., Thabane, L., Anis, A.H. (2018). Output from the CIHR Canadian HIV Trials Network international postdoctoral fellowship for capacity building in HIV clinical trials. *HIV/AIDS – Research and Palliative Care*, 10, 151–155.

McKay. D.M. & Daniels, S. (2003). Canadian Association of Gastroenterology – Canadian Institutes of Health Research – pharmaceutical partner postdoctoral operating fellowship programme: An outstanding success that continues to excel! *Canadian Journal of Gastroenterology*, 17, 437–439.

Miller, J.M & Feldman, M.P. (2014). The sorcerer's postdoc apprentice: uncertain funding and contingent highly skilled labour. *Cambridge Journal of Regions, Economy and Society*, 7, 289–305.

Mongeon, P. & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. Scientometrics, 106, 213–228.

Nowell, L., Ovie, G., Berenson, C., Kenny, N. & Hayden, K.A. (2018). Professional learning and development of postdoctoral scholars: A systematic review of the literature. *Education Research International*, 2018, Article ID 5950739.

Nowell, L., Ovie, G., Kenny, N. & Hayden, K.A. & Jacobsen, M. (2019). Professional learning and development initiatives for postdoctoral scholars. *Studies in Graduate and Postdoctoral Education*, 11, 35–55.

Numminen, O., Virtanen, H., Hafsteinsdóttir, T. & Leino-Kilpi, H. (2020). Postdoctoral nursing researcher career: A scoping review of required competences. *Nursing Open*, 7, 7–29.

Pepper, J., Krupińska, O.D., Stassun, K.G. & Gelino, D.M. (2019). What does a successful postdoctoral fellowship publication record look like? *Publications of the Astronomical Society of the Pacific*, 131, Article 014501.

Peters, M.D.J., Godfrey, C., McInerney, P., Munn, Z., Tricco, A.C. & Khalil, H. (2020). Chapter 11: Scoping reviews (2020 version). In E. Aromataris & Z. Munn (Eds.), *JBI Manual for Evidence Synthesis* (pp. 406–451). Joanna Briggs Institute, University of Adelaide.

Pion, G. & Ionescu-Pioggia, M. (2003). Bridging postdoctoral training and a faculty position: Initial outcomes of the Burroughs Wellcome Fund Career Awards in the Biomedical Sciences. *Academic Medicine*, 78, 177–186.

Ranieri, V., Barratt, H., Fulop, N. & Rees, G. (2016). Factors that influence career progression among postdoctoral clinical academics: A scoping review of the literature. *BMJ Open*, 6, e013523.

Reskin, B.F. (1976). Sex-differences in status attainment in science: The case of the postdoctoral fellowship. *American Sociological Review*, 41, 597–612.

Schneider, J.W. & van Leeuwen, T.N. (2014). Analysing robustness and uncertainty levels of bibliometric performance statistics supporting science policy: A case study evaluating Danish postdoctoral funding. *Research Evaluation*, 23, 285–297.

Sherry, D., Fennessy, M.M., Benavente, V.G., Ruppar, T.M. & Collins, E.G. (2013). Important considerations when applying for a postdoctoral fellowship. *Journal of Nursing Scholarship*, 45, 210–218.

Thompson, D.F. & Walker, C.K. (2015). A descriptive and historical review of bibliometrics with applications to medical sciences. *Pharmacotherapy*, 35, 551–559.

Treichler, E.B.H., Crawford, J.N., Higdon, A. & Backhaus, A.L. (2020). Diversity and social justice training at the postdoctoral level: A scoping study and pilot of a self-assessment. *Training and Education in Professional Psychology*, 14, 126–137.

van Leeuwen, T. (2004). Descriptive versus evaluative bibliometrics. In H.F. Moed, W. Glänzel & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research*, pp. 373–388. Dordrecht: Springer.

Zabetta, M.C. & Geuna, A. (2019). International postdoctoral mobility and career effect in Italian academia: 1986–2015. In G. Catalano, C. Daraio, M. Gregori, H.F. Moed & G. Ruocco (Eds.), *Proceedings of the 17th International Conference on Scientometrics & Informetrics* (ISSI 2019), (Volume II, pp. 2448–2459). Rome: Edizioni Efesto.

Zhu, J. & Liu, W. (2020). A tale of two databases: The use of Web of Science and Scopus in academic papers. *Scientometrics*, 123, 321–335.

Zumeta, W. (1985). *Extending the Educational Ladder: The Changing Quality and Value of Postdoctoral Study*. Massachsetts: Lexington.

Why university affiliation discrepancies still exist in citation indexes and how to resolve them

Philip J. Purnell[1,2]

p.j.purnell@cwts.leidenuniv.nl
[1]Centre for Science and Technology Studies,
Leiden University, P.O. Box 905, 2300 AX Leiden, The Netherlands
[2]United Arab Emirates University, Al Ain, UAE

Abstract

Research managers benchmarking universities against international peers face the problem of affiliation disambiguation. Traditional bibliometric data sources such as Scopus and Web of Science each conduct a manual process to describe organization relationships between academic institutions which allow the user to search a single unified name. This study used digital object identifiers (DOIs) as the common attribute for publications indexed in four bibliometric databases to retrieve publications from 18 Arab universities. We then determined the overlapping coverage and identified discrepancies between databases that were due to different ways of treating affiliation names. We found the larger databases, Dimensions and Microsoft Academic tended to have more affiliation discrepancies with each other and with the more selective Web of Science and Scopus. Finally, we manually examined sample records that revealed the reason for the affiliation discrepancies. There was an even spread between missing affiliations, unification differences, and assignation to the wrong institution. We conclude by calling for renewed efforts to develop a global, open-source disambiguation system which should be maintained by universities and incorporated into all bibliometric data sources.

Introduction

The problem of affiliations

The research community understands to varying degrees the importance of getting its university affiliation names right. Individual researchers are now routinely assessed at least in part on their ability to produce published articles and their institutions are at least partially ranked on those very same papers. The single factor tying the paper to its author's employer is the affiliation name given by the author when they submit the manuscript to a journal.

There are many ways of acknowledging a university and one can easily confuse a rudimentary database by simply swapping My City University with University of My City. Other common variations involve acronyms, MCU, UMC, or partial acronyms, MC University, Univ MC. The list easily extends to dozens of variants when authors introduce their faculty or department name sometimes at the expense of the university name. Add to that the common practice of incorporating one institution into another or splitting part of a university away from its main organization, along with larger mergers and creation of international branch campuses and we have a complex problem for those assessing the university.

Indeed, nowadays journal and authors names are relatively constant while it is not uncommon for university names to change. Although there have been several initiatives to address the problem by using unique identifiers for research institutions, none have been universally adopted to the same extent as for journals (ISSN), individuals (ORCID), or documents (DOI). These efforts have mainly been made by the major citation indexes such as Scopus (Affiliation Identifier or AFID), Web of Science (Organization Enhanced), Dimensions (Global Research Identifier Database or GRID). A new community-led collaboration of multiple organizations has been launched and is known as the Research Organization Registry (ROR). It holds promise because it is closely linked to GRID and is to be incorporated into Crossref metadata (Lammey Rachael, 2020).

Databases used in such assessments have made strides into resolving this problem using different solutions including manual submission of affiliation variant lists by universities to database owners or automated unification systems. The degree of accuracy is still unquantified and policy makers who rely on bibliometric analysis often overlook an inherent level of error when using these data sources.

The importance of affiliations

Many universities driven by increased external competition (Brankovic, Ringel, & Werron, 2018; Espeland & Sauder, 2016) seek to maximize their position in the various international ranking tables. The ranking organizations in turn typically assess institutions' performance against a set of criteria that usually include the quantity and impact of research publications (Centre for Science & Technology Studies Leiden University, 2020; QS Intelligence Unit, 2019; Shanghai Ranking Consultancy, 2019; Times Higher Education, 2019; US News & World Report LP, 2019). These ranking systems use either Elsevier's Scopus or Clarivate Analytics' Web of Science (Web of Science) to compute the bibliometric component of their tables.

The level of accuracy of those databases and their ability to assign papers to the correct affiliations consequently becomes one of the limiting factors in an institution's performance (Orduna-Malea, Aytac, & Tran, 2019). Any 'missing' papers can cost a university valuable places in the ranking table and authors are routinely encouraged to use the official institution name when publishing. Nevertheless, each year universities complain to the ranking systems of missing papers, but the rankers are constrained by the limitations of the citation indexes. Disgruntled universities are usually referred to the database owners to resolve their affiliation-related complaints.

This study focused on those records found in overlapping database coverage but with discrepancies in the university affiliation name between databases. We manually examined two dozen records from each surplus and attempted to explain the mismatch in the context of affiliation indexing. This has major implications for any university benchmarking study and particularly the international university rankings. We chose to make this a regional study because it requires close knowledge of university names, and we selected 18 universities from the Arab region because of our familiarity with this region.

Affiliation disambiguation policy

In Scopus, we used the Affiliation identifier (AFID) which is a unique identifier for the institution to which records are tagged. In many cases however, Scopus includes multiple AFIDs for a single university treating medical schools and other faculties as separate entities. This does not appear to be consistent and for some universities there is only one AFID while for others there are many. Although there is a search option that unifies the AFIDs for one single institution, some of the component organizations appear less related to the main institution than others and we therefore used only the main AFID for each university. From Web of Science, we used organization enhanced (OE) which is a preferred institutional name searchable in the database and to which records from that organization and its component parts are unified. The OE unification process is performed by the database owner with voluntary input from the institutions themselves. Dimensions and Microsoft Academic each use GRID which was developed by Digital Science to describe both parent-child relationships between institutions and external related organizations. GRID disambiguates affiliation names for approximately 100,000 organizations and we therefore used the GRID record linked to the generally accepted name for each of the 18 university names.

Research questions

1. How prevalent are affiliation discrepancies between citation indexes for Arab universities?

2. What causes affiliation discrepancies and what are the implications for databases and bibliometricians?

The answers to these questions will be useful in our understanding of the extent to which research outputs are covered by the different databases. Since many decisions are based on the outcome of bibliometric studies, comparisons, and university rankings, policy makers will be better informed about the limitations of bibliometrics studies and comparisons. The ranking bodies may take these limitations into account when they publish their league tables. Database owners may incorporate these finding into their development plans and algorithms to improve the accuracy of their products and make them more competitive.

Literature review

Work on author affiliations was documented at Leiden University in the mid-1980's. Indeed the LISBON Institute, the predecessor of CWTS already used affiliation data from the Science Citation Index (SCI) to report the changes in academic collaboration in the Arab region following the geopolitical developments in the 1980s (DeBruin, Braam, & Moed, 1991). The problems of author affiliations became more important as bibliometric reports gained popularity and started being offered as a service (Calero-Medina, Noyons, Visser, & De Bruin, 2020) and the development of in-house citation indexes at CWTS in the 1990s were constructed in a manner that facilitated address disambiguation.

Interesting applications of author affiliations included delimitation of scientific subfields using terms in corporate addresses in scientific publications and assessing whether European funding had spurred increased co-operation between EU Member States using the country name in the author address fields. The problem of missing author affiliations has been largely addressed but still in 2015 the Web of Science indexes contain sizeable quantities of publications without any affiliations whatsoever (SCIE: 7.6%, SSCI: 6%, and A&HCI: 35%) (Liu, Hu, & Tang, 2018).

The Web of Science OE feature attempts to overcome the affiliation disambiguation problem and has been selectively used in bibliometric studies to improve accuracy (Baudoin, Akiki, Magnan, & Devos, 2018). A recent study (Donner, Rimmert, & van Eck, 2020) showed widely varying recall and precision across institutions between Web of Science OE, Scopus AFID and a German institution affiliation disambiguation system described as 'near-complete' for German public research organizations. Taking the German system as ground truth, the authors concluded the resulting inconsistencies in publication and citation indicators using the commercial vendor systems should be taken into consideration by policy makers.

One of the first large-scale studies comparing database coverage across a multi-institution dataset between Scopus, Web of Science, and Microsoft Academic examined the publication overlap of 15 universities with DOIs serving as the common attribute (Huang et al., 2020). The authors created a Venn diagram for each university showing the proportion of DOIs indexed by all three data sources. The diagram also revealed the extent to which DOIs were covered by only one of the three databases, that we term a surplus. For instance, DOIs found in Web of Science but not in Scopus or Microsoft Academic count as Web of Science surplus.

Method

This study examines the overlap of indexed DOIs from for 18 universities (Table 1) of all document types, in four international, multidisciplinary citation indexes often used in bibliometric studies, namely Scopus, Web of Science, Dimensions, and Microsoft Academic. From Scopus we used the main affiliation ID only because of the differences in the level of disambiguation performed for the universities. Within the Web of Science, we used five editions; the Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, the Conference Proceedings Citation Index-Science, and the Conference Proceedings Citation Index-Social Sciences & Humanities. Neither the Book

Citation Index (BkCI), nor the Emerging Sources Citation Index (ESCI) were used because we do not have access to them. We refer to our version of this database henceforth as CWTS Web of Science. From Dimensions, we extracted only publications because these are comparable with the documents in the other databases, but not grants, patents, datasets, clinical trials, or policy documents. We did not use Google Scholar because of the difficulties in accessing its data, especially for large bibliometric studies. All data were retrieved from the versions of the databases available in the database system of the Centre for Science and Technology Studies (CWTS) at Leiden University. The data was received in March 2020 (CWTS Web of Science), April 2020 (Scopus and Dimensions), and July 2020 (Microsoft Academic).

Table 1: The universities used along with their abbreviations

Abbreviated name	*Full institutional name*	*Country*
Ain Shams	Ain Shams University	Egypt
Alexandria	Alexandria University	Egypt
Assiut	Assiut University	Egypt
AUB	American University of Beirut	Lebanon
Babylon	University of Babylon	Iraq
Baghdad	University of Baghdad	Iraq
Bahrain	University of Bahrain	Bahrain
Carthage	University of Carthage	Tunisia
Jordan	University of Jordan	Jordan
Khalifa	Khalifa University	United Arab Emirates
King Abdulaziz	King Abdulaziz University	Saudi Arabia
King Saud	King Saud University	Saudi Arabia
Kuwait	Kuwait University	Kuwait
Qatar	Qatar University	Qatar
Lebanese	Lebanese University	Lebanon
Sfax	University of Sfax	Tunisia
Sultan Qaboos	Sultan Qaboos University	Oman
UAEU	United Arab Emirates University	United Arab Emirates

We subsequently manually analyzed 24 samples of affiliation discrepancies from each of the database pairs. For each pair we selected the university with the highest proportion of affiliation discrepancy provided that university had not already been used. If so, we moved on to the university with the next highest proportion. Examination was performed by manually searching for the records on the web interface versions of each database and checking the published PDF documents as the ground truth.

Results and discussion

Differences between databases

It was not possible to approach this study from the overlapping coverage between databases for any given university. Overlapping coverage would necessarily contain the correct affiliation name in each database, otherwise it would not be retrieved. Instead, we searched for DOIs containing affiliations from each of the selected universities in one database that we referred to

as the primary database. Then we searched the comparator databases to determine first, whether the DOI was covered, and second whether it was assigned the same affiliation name as in the primary database. Consequently, for each database pair we present our results as a series of stacked bars. At the ends of each bar are those records with affiliation discrepancies between the two databases while overlapping affiliations are found in the central portion of the bar.
In Table 2 we present the overall overlapping DOI count for a cumulative dataset of the 18 selected universities for each pair of databases. The right-hand portion of the bar represents those DOIs found in the primary database with one of the university affiliations that is also present in the comparator database but not assigned to the same affiliation. Without closer examination of the individual records, we cannot say whether the affiliations are correct or erroneous in either of the databases. Also, if an affiliation is incorrect in both databases, then it would be missed from the entire dataset.

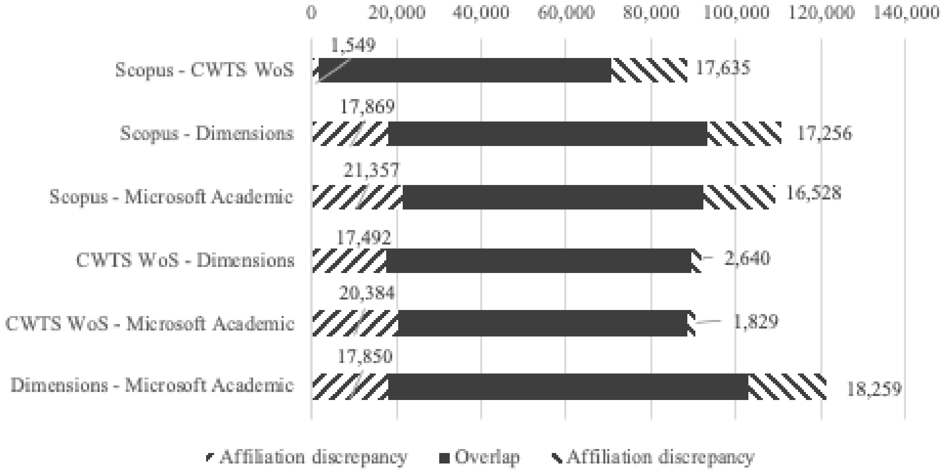


Figure 1. Affiliation discrepancies by database pair

The first bar shows that 1,549 CWTS Web of Science records for the 18 universities were also found to be covered in Scopus but not with the same affiliations. This is a small proportion of the overall overlapping DOIs found when compared with the proportion of Scopus records found to be present in CWTS Web of Science but with affiliation discrepancies. The fourth and fifth bars show a similar pattern with a small proportion of CWTS Web of Science papers showing affiliation discrepancies with either Dimensions or Microsoft Academic. The third bar shows a larger proportion of Microsoft Academic records that do not agree with their corresponding record in Scopus than there are Scopus records with affiliation discrepancies in Microsoft Academic.
These differences might be explained by CWTS Web of Science and to a lesser extent Scopus having more accurate affiliation data than Dimensions and Microsoft Academic resulting in fewer discrepancies. Alternatively, the differences could be related to the extent and type of sources these databases cover as that might impact their likelihood to have DOIs or complete and accurate affiliation data. For instance, CWTS Web of Science is only a subset of the full Web of Science database. Our version of the database excludes the BKCI and its component document types, books and book chapters which frequently lack a DOI (Gorraiz, Melero-Fuentes, Gumpenberger, & Valderrama-Zurián, 2016). It also excludes the ESCI that comprises several thousand regional or peripheral journals that might well be present in the larger databases but less frequently be associated with DOIs. As only those records with DOIs are

included in the study, there is increased likelihood of finding them in CWTS Web of Science. Indeed, Web of Science and Scopus are smaller and more selective databases and consequently more likely to comprise journals articles from publishers that register DOIs for their articles. Any article in a comparator database without a link to its DOI would be missed in this study. Dimensions and Microsoft Academic are both considerably larger than Web of Science and Scopus and divergence is therefore more, rather than less likely. Two recent large-scale comparisons of these data sources reveal closer overlap between Scopus and Web of Science than between Scopus and Microsoft Academic (Huang et al., 2020; Visser, Van Eck, & Waltman, 2021). The Visser study also showed Scopus to hold closer overlap with Web of Science than with Dimensions. The leaves us with the probability that affiliation discrepancies are greater between Scopus on the one hand and Dimensions and Microsoft Academic on the other than they are between Scopus and Web of Science.

Reasons for mismatches on affiliations

We manually examined two dozen records from each of the database pairs using the web interface for each database. For each of these examples, we attempted to discover the reason the affiliation was responsible for the DOI not being retrieved in the search. The main reasons are summarized in table 2.

Table 2. Affiliation discrepancies

Database pair	*Institution*	*Missing affiliation*	*Missing second affiliation*	*Assigned to wrong institution*	*Unification*	*Inconclusive*	*Total*
Scopus-CWTS WoS	Lebanese			1	22	1	24
Scopus-Dimensions	Sfax	11			11	2	24
Scopus-Microsoft Academic	Khalifa	2			20	2	24
CWTS WoS - Scopus	Saud	6	1	13		4	24
CWTS WoS - Dimensions	AUB	6	1	1	14	2	24
CWTS WoS - Microsoft Academic	Assiut	6	12	5	1		24
Dimensions - Scopus	Carthage	3			21		24
Dimensions - CWTS WoS	Bahrain	1		19	4		24
Dimensions - Microsoft Academic	Babylon	9	6	7	2		24
Microsoft Academic - Scopus	Kuwait					24	24
Microsoft Academic - CWTS WoS	UAEU	4	1	3	15	1	24
Microsoft Academic - Dimensions	Qatar	20		2	2		24

Missing affiliation

The author affiliation has not been captured by the database and therefore a search for the affiliation name did not retrieve the record. We found almost all (20 of the 24) Qatar University affiliation discrepancies in the Microsoft Academic–Dimensions pair were caused by missing affiliations in Dimensions. In most of these cases all author affiliations were missing, and the papers were conference proceedings or book chapters. Similarly, we found where CWTS Web of Science has missed author affiliations, the majority were meeting abstracts.

Missing second affiliation

The author's first affiliation has been listed but not the second. This is similar to the above category but worth separating as it appears that Microsoft Academic has distinct groups of papers for which the additional affiliation is missed while the first is captured. As an example, on 10.1166/asl.2017.7424 the PDF shows University of Babylon as second affiliation for one author that is omitted from the record in Dimensions and Microsoft Academic, but included in Scopus and Web of Science. In other cases (e.g., 10.1016/j.asoc.2016.06.019), the second affiliation is listed separately from the first on the PDF under categories such as 'Author's current address' or in this case, 'Correspondence address'. Scopus and Web of Science included this as second affiliation, while Dimensions and Microsoft Academic did not.

Assigned to the wrong institution

We found seven records from University of Babylon erroneously assigned in Microsoft Academic to either the College of Information Technology or the Information Technology University in Pakistan, or the University College of Engineering in India. Each of these are in fact sub-units of the University of Babylon. Lots of universities have a College of Information Technology and we found several other examples of papers from e.g. College of Information Technology, UAEU that were also assigned to the College of Information Technology in Pakistan. Due to erroneous assignments the publication counts for these universities will be too low and the publication count for the College of Information Technology in Pakistan will be too high.

Other examples showed records assigned to UAEU in Web of Science OE were in fact published by authors at a Moroccan institution called Université Abdelmalek Essaadi sometimes abbreviated to UAE University and mistakenly unified to the wrong institution.

Similarly, authors from LaSTRe Laboratory in Tripoli, Northern Lebanon, which is affiliated to the Lebanese University in Beirut, had been erroneously affiliated to the University of Benghazi in Tripoli, Libya in the Web of Science. Two further records from the same university were assigned to the United States of America due to confusion over the tiny town of Lebanon in Grafton County, New Hampshire. It appears therefore that the presence of a city name or country name might sometimes trigger unification to the wrong organization enhanced name in the Web of Science.

Unification

An affiliation is listed but it has not been unified to the main or correct university. For example, 10.1002/2015gl066534 features the affiliation Masdar Institute which has been unified to Khalifa University in Dimensions but not in Microsoft Academic. Khalifa University is the result of a merger between three institutions, the Petroleum Institute, Masdar Institute and Khalifa University of Science and Technology in 2017. Microsoft Academic's approach is not necessarily wrong because Masdar Institute existed when the article was published but highlights decisions that have to be made when handling organization mergers.

Many authors in Tunisia have acknowledged their institution as Faculty of Sciences at Sfax. We found that Scopus unified these papers to the University of Sfax while Dimensions treated it as a separate organization. In a similar case, 10.1002/ajh.25075 the American University of Beirut Medical Centre has correctly been unified to AUB in Web of Science but not in Dimensions.

Also in Tunisia, we found several records affiliated to the National Institute of the Applied Sciences and Technology and several other research centers which Web of Science and Dimensions unified to the University of Carthage while Scopus did not. In most of these cases, the university was not mentioned on the PDF, but Scopus has made the link through its disambiguation process.

Inconclusive

We classified as inconclusive cases where there was no obvious reason for the DOI not being retrieved, a human indexing decision involved, or access to PDF proved impossible. An example of a human indexing decision is a book preface with a DOI (10.1016/B978-0-12-800887-4.00034-1) but no authors or affiliations. While Scopus had assigned the book editors as authors, Dimensions had not. Another interesting case was a letter to the editor published by three authors with a long list of additional signatories at the end. Scopus had counted all the signatories as authors of the letter while Dimensions limited authorship to the three at the top of the paper. Neither of these cases is clear cut, if one accepts the book editors in the first case, and letter signatories in the second should be named as authors then their affiliations are missing in Dimensions. If they should not be named, then they are phantom affiliations in Scopus.

Conclusions

Each database uses a different method of disambiguation and unification. That makes comparison difficult. Scopus AFIDs and Web of Science Organization Enhanced records are very useful because they allow the user to analyze the whole university or any of its components and affiliated institutions. When comparing institutions, we make the assumption the disambiguation process has been performed to the same level for all institutions in the analysis. Our results show this is not always true and that some comparisons will produce misleading results. In Scopus, the relationship between the 'Affiliation only' AFID and those included in the 'Whole institution' appear to vary widely in their level of unification. In Web of Science, there is evidence of errors brought into affiliation assignation by the presence of certain city and even country names. Dimensions has missing affiliations in some conference proceedings paper and book chapters while Microsoft Academic has mistakenly affiliated many records to the wrong institution.

A considerable limitation to the present and most previous studies on affiliation disambiguation is the fact that university names change over time. Changes result from a number of factors including mergers and splits but also for reasons of government naming conventions, changes of country leaders, city names, and other factors. Once an address is unified to an affiliation, all its prior papers will be found when searching the new unified affiliation. Studies are therefore a time-frozen shot of the current unification and do not take account of unification dynamics over time.

Some university rankings providers assume responsibility for disambiguating affiliations in their league tables to different degrees. This is especially true in the case of the Leiden Ranking which disambiguates all affiliations from its proprietary database (Calero-Medina et al., 2020), while both QS (QSIU, 2019) and Times Higher Education have begun supplementary work on Scopus unification in special cases. This practice is welcomed and providers of products that derive from bibliometric data sources should be encouraged to increasingly take stewardship of the analytical process and responsibility for the accuracy of the resulting publication.

There is a need for a universal unique identifier for academic institutions that should reflect the current and historical organization relationship tree. The ideal indicator will be supported by input from the institutions themselves in the same way that researchers maintain their own ORCID records. That way the accuracy, maintenance, and historical record will be maximized. There will be scope for non-maintenance or misuse especially where an institution can benefit from a certain interpretation of its organization, but these will be outweighed by the benefits. Universities and their stakeholders should still decide their own names and they are still the most appropriate managers of the public record of their relationships with sub-units and external entities. However, a global initiative is required to provide an integrated platform for such relationships to be presented. The Research Organization Registry is a promising initiative that will fill that need especially if adopted by all data source providers.

Acknowledgements

Ton van Raan and Ludo Waltman for expert guidance throughout the study, input on the article, and insights into the Leiden Ranking, Ashwin Fernandes, Duncan Ross for advice on university ranking disambiguation process.

References

Baudoin, L., Akiki, V., Magnan, A., & Devos, P. (2018). Production scientifique des CHU-CHR en 2006–2015 : évolutions et positionnement national. *La Presse Médicale*, *47*(11, Part 1), e175–e186. https://doi.org/https://doi.org/10.1016/j.lpm.2018.06.016

Brankovic, J., Ringel, L., & Werron, T. (2018). How rankings produce competition: The case of global university rankings. *Zeitschrift Fur Soziologie*, *47*(4), 270–288. https://doi.org/10.1515/zfsoz-2018-0118

Calero-Medina, C., Noyons, E., Visser, M., & De Bruin, R. (2020). Delineating Organizations at CWTS---A Story of Many Pathways. In C. Daraio & W. Glänzel (Eds.), *Evaluative Informetrics: The Art of Metrics-Based Research Assessment : Festschrift in Honour of Henk F. Moed* (pp. 163–177). https://doi.org/10.1007/978-3-030-47665-6_7

Centre for Science & Technology Studies Leiden University. (2020). Indicators. Retrieved August 11, 2020, from https://www.leidenranking.com/information/indicators

DeBruin, R. E., Braam, R. R., & Moed, H. F. (1991). BIBLIOMETRIC LINES IN THE SAND. *NATURE*, *349*(6310), 559–562. https://doi.org/10.1038/349559a0

Donner, P., Rimmert, C., & van Eck, N. J. (2020). Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quantitative Science Studies*, *1*(1), 150–170. https://doi.org/10.1162/qss_a_00013

Espeland, W. N., & Sauder, M. (2016). Engines of anxiety: Academic rankings, reputation, and accountability. In *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85011865409&partnerID=40&md5=f6d9c54cf733b6b09f082abcdfe745e1

Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., & Valderrama-Zurián, J. C. (2016). Availability of digital object identifiers (DOIs) in Web of Science and Scopus. *Journal of Informetrics*, *10*(1), 98–109. https://doi.org/10.1016/j.joi.2015.11.008

Huang, C.-K. (Karl), Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., & Ozaygen, A. (2020). Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies*, *1*(2), 445–478. https://doi.org/10.1162/qss_a_00031

Lammey Rachael. (2020). Solutions for identification problems: a look at the Research Organization Registry. *Sci Ed*, *7*(1), 65–69. https://doi.org/10.6087/kcse.192

Liu, W., Hu, G., & Tang, L. (2018). Missing author address information in Web of Science - An explorative study. *Journal of Informetrics*, *12*(3), 985–997. https://doi.org/10.1016/j.joi.2018.07.008

Orduna-Malea, E., Aytac, S., & Tran, C. Y. (2019). Universities through the eyes of bibliographic databases: a retroactive growth comparison of Google Scholar, Scopus and Web of Science. *Scientometrics*, *121*(1), 433–450. https://doi.org/10.1007/s11192-019-03208-7

QS Intelligence Unit. (2019). QS World University Rankings. Retrieved August 11, 2020, from http://www.iu.qs.com/university-rankings/world-university-rankings

QSIU. (2019). Papers & Citations. Retrieved December 10, 2020, from QS Intelligence Unit website: http://www.iu.qs.com/university-rankings/indicator-papers-citations/

Shanghai Ranking Consultancy. (2019). Academic Ranking of World Universities Methodology. Retrieved August 11, 2020, from http://www.shanghairanking.com/ARWU-Methodology-2019.html

Times Higher Education. (2019). THE World University Rankings 2020: methodology.

US News & World Report LP. (2019). How U.S. News Calculated the Best Global Universities Rankings. Retrieved August 11, 2020, from https://www.usnews.com/education/best-global-universities/articles/methodology

Visser, M., Van Eck, J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *ArXiv*. Retrieved from https://arxiv.org/abs/2005.10732

Media reporting of cancer research in China and the UK, 2014-19

Yuping Qu[1], Elizabeth Davies[2] and Grant Lewison[1]

[1] *yuping.qu@kcl.ac.uk, grantlewison@aol.co.uk*
King's College London, Department of Cancer Policy and Global Health, Guy's Hospital, Great Maze Pond, London SE1 9RT (UK)

[2] *elizabeth.davies@kcl.ac.uk*
King's College London, Department of Cancer Epidemiology, Population and Global Health, Guy's Hospital, Great Maze Pond, London SE1 9RT (UK)

Abstract

This paper reports research in progress on a PhD project to investigate how cancer research is reported in some newspapers in China and the UK during the six years 2014-19. The stories were selected if they cited identifiable cancer research papers recorded by the Web of Science, whose details were appended to those of the news stories in a big spreadsheet. We are comparing the coverage by cancer site (*e.g*., lung, breast) with the disease burden from different cancers in the two countries, and also with the amount of research taking place. We are also examining the location of the researchers whose work is being cited to see how much own-country work is being described, and whether research outwith the main cities is being neglected. Many research stories also quote commentators who can put the results in context. Finally, we are planning to determine if citation of research in newspapers leads to more than expected numbers of citations from researchers in the country of their publication. Some early results from the popular Chinese newspaper, *Global Times*, and *New Scientist,* are presented.

Introduction

Cancer is now one of the leading causes of death world-wide and its burden in the population has increased over the last 30 years, see Figure 1. This is because major improvements in infant and maternal mortality, and reductions in many infectious diseases have combined to increase life expectancy. Consequently, more people survive into old age where they are more susceptible to cancer.

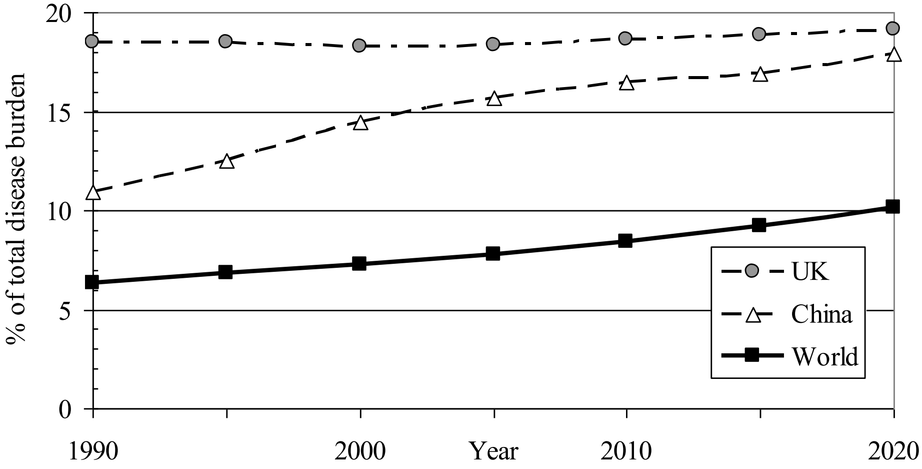


Figure 1. Cancer burden as a percentage of the total, Disability-Adjusted Life Years (DALYs), for the world, China, and the UK, from 1990 to 2020. ***Data from the Institute for Health Metrics and Evaluation at the University of Washington (GBD Compare | IHME Viz Hub (healthdata.org)).***

As China has developed economically over this period, its population has become more susceptible to cancer, and its relative cancer burden has increased and is now close to that of the UK. It is therefore as major a concern for the Chinese government as it is in the UK.

The causes of cancer are numerous, and over a thousand have been listed by the International Agency for Research on Cancer (IARC), part of the World Health Organization Table 4 (who.int). Genetics plays a large part, and some of these environmental causes are hard for people to avoid, but there are several causes of cancer that are to a large extent under the control of individuals. These include whether to smoke or drink alcohol, what to eat, and how much exercise to take. There are public health campaigns on each of these in many countries, some sponsored by the government, and in the UK, some by charities. However, one of the main means by which people obtain health information is from the mass media.

This project, therefore, seeks to assess how information about cancer research is reported and presented to the public in two very different countries, China and the UK. They are different because the media, particularly newspapers, are controlled in very different ways: by the communist central government in China, and by their capitalist proprietors in the UK. [An exception which we are studying is *The Guardian*, which is owned by a private-non-profit organisation, the Scott Trust.] These countries also differ because the distribution of cancers differs between the two countries, see Figure 2.

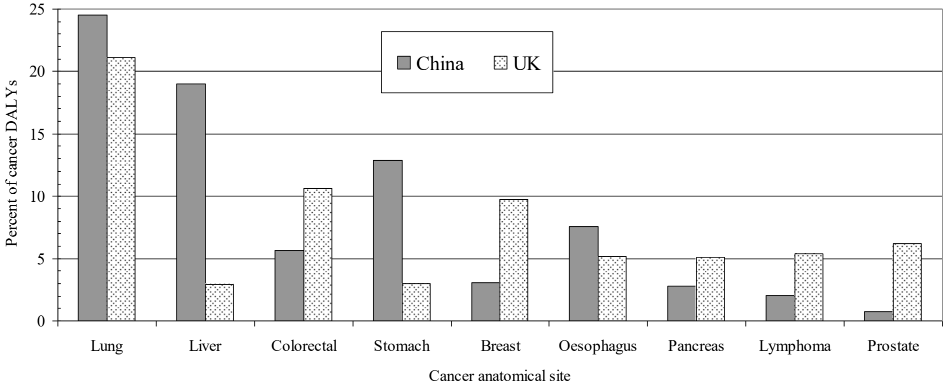


Figure 2. The percentages of total cancer DALYs in 2015 from different cancers in China and the UK. Data from the World Health Organization (Indicators (who.int))

Lung cancer is the cancer with the greatest burden in both countries, while liver, stomach and to some extent oesophageal cancers are more important in China. Colorectal, breast and prostate cancers also contribute to relatively larger burdens in the UK. One of the objectives of our research is to assess whether media reports take account of and reflect these different epidemiological patterns.

Both China and the UK are now very active in cancer research. China's research output, measured by papers in the Web of Science (WoS), exceeded that of the UK in 2009 (Figure 3), and that of the USA in 2018. [The slowdown since 2015 is probably an artefact caused by the addition of the Expanded Sciences Citation Index to the WoS in that year.] So we will also be investigating whether the newspapers in each country choose to cite preferentially cancer research from their own countries, relative to research from the world total, and whether they tend to favour regions of each country where research is concentrated. These regions include the big cities of the eastern provinces in China, especially Beijing and Shanghai (Li & Lewison, 2020), and the "Golden Triangle" of Cambridge, London and Oxford in the UK (Begum *et al*., 2017).

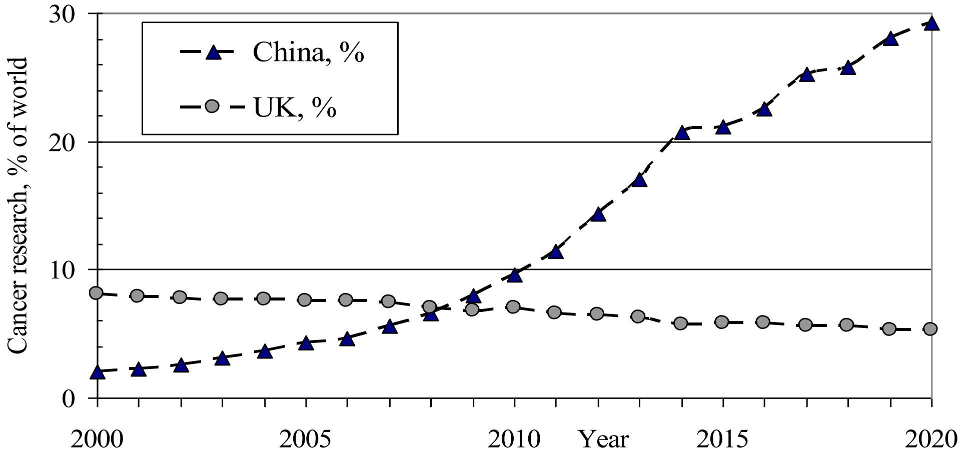


Figure 3. Cancer research outputs of China and the UK in the WoS, 2000-20.

Media reports of medical research have become somewhat polarised recently, with many false stories appearing, notably about the COVID-19 vaccines. We therefore consider it important to check whether journalists writing articles are seeking outside independent commentators who can put the research in context, and verify that it is good science. Previous studies of the UK media (Lewison *et al.*, 2008; Pallari *et al.,* 2020) have shown that British journalists frequently quote external experts, particularly from the cancer charities, to validate the results that they are reporting. One question is therefore, Do Chinese journalists use external experts, and if so, where do they find them? The tone of the stories is also of concern, as (at least in the UK) newspapers often over-hype their stories so as to grab the attention of their readers and increase circulation. Our question for China is, Does the Chinese Communist Party (CCP) control of the media provide an opportunity to demonstrate to its citizens the superiority of Chinese cancer research?

A final investigation will be to assess whether reportage in the mass media has any measurable influence. It is only one of the many factors that can influence health-related behaviour, but it has been shown that news reports can lead to more citations in journals (Phillips *et al.*, 1991; Lewison *et al.*, 2008). In the absence of an unusual "before and after" scenario such as the one Phillips investigated when *The New York Times* employees went on strike and then resumed work, it will be necessary to see if reportage in a country's media increases the proportion of citations to the cited papers that are from that country. There are, of course, many other possible effects of media reports on people's behaviour, but these are much harder to demonstrate.

Methodology

The first task was to select the media for study. In China, we chose eight mass media, seven of which are newspapers, and one popular online medical daily, *Dr. Ding Xiang*. They are normally read in their online versions with printed editions available (see Table 1). The newspapers are all controlled by the Chinese Communist Party (CCP) so they are expected to reflect the views of the Government. For example, the *People's Daily* is the largest newspaper in China, and the most important wind vane for the party's theory, line, principles, and policies (People's Daily, 2017). But some, such as the *South China Morning Post* and the *Beijing News,* try to maintain a certain degree of independence. *Global Times* is very widely read, and its original intention was to present mainly foreign news with a large network of correspondents from all over the world (Global Times, 2019). Currently it has become the key platform to cultivate and shape the popular consensus according to China's will (Lee, 2010). Three of the newspapers, the *Beijing News, Xin Min Evening News* and *Heilongjiang Daily*, are regional, so might be expected to take the local cancer burden and the regional differences in potentially modifiable cancer risks (Chen *et al*., 2019) into account. *Dr Ding Xiang* is the only Chinese medium we selected without a printed version. It is the most popular and influential WeChat Public Account for health in China now, with professional contributors, and more than 10 million followers. It also explicitly aims to fight health misinformation, such as that garlic can prevent cancer, and to bridge grassroots advice and that from medical experts. (Wang, 2018).

Table 1. List of news media in China selected for study

Title (CN)	*Title (English)*	*Area*	*Language*
人民日报	*People's Daily*	Mainland China	Simplified Chinese
环球时报	*Global Times*	Mainland China	Simplified Chinese
中国日报	*China Daily*	Mainland China	English
新京报	*Beijing News*	Mainland China	Simplified Chinese
南华早报	*South China Morning Post*	Hong Kong	English
丁香医生	*Dr Ding Xiang*	Mainland China	Simplified Chinese
新民晚报	*Xin Min Evening News*	Shanghai	Simplified Chinese
黑龙江日报	*Heilongjiang Daily*	Heilongjiang	Simplified Chinese

[The Chinese written language was simplified in 1949, but the original script is still used in Hong Kong and Taiwan.]

In the UK, where the news media are independent and the market is segmented by education level and politics, we selected ones that represented a wide range of views, shown in Table 2. Data were already available for *Daily Mail* and *The Guardian* for the years 2002-13, and for *New Scientist* for some of the study period, and for five years before then. The others were chosen to represent an intellectual and financial readership, one that is "middle-of-the-road", and one that is distinctly popular and un-intellectual.

Table 2. List of news media in the UK that were selected for study

Title	*Circulation*	*Characteristics*
Daily Mail	1,134,184	Right-wing tabloid, middle-market readership
Financial Times	155,009	Centre-right business broadsheet, international readers
The Guardian	111,155	Left-wing Berliner format, intellectual
i	215,932	Centre-left tabloid, parent is online only
The Sun	1,206,595	Right-wing populist tabloid, low education level readers
New Scientist	118,009	Centre-left popular weekly science magazine

The individual stories were first identified through their headlines and brief synopses on the Factiva database by means of a simple search statement:

> *(cancer OR leukaemia* OR melanoma* OR lymphoma* OR tumour* OR sarcoma*) AND (research* OR study OR scientist* OR expert*)*

and these details were downloaded to an Excel spreadsheet. They were first formatted so that all the data for each story (including date, the name of the journalist, and length in words) were on a single line. Then they were filtered to remove any that did not have one of some 16 words in their headline or synopsis indicating that they cited published research. The full texts of those that remained were then individually downloaded *via* Factiva and read. Further details were added to the spreadsheet, including those of the cited research paper (author, institution, journal) and of any commentator(s).

The next stage was to identify each cited paper in the WoS and download all its bibliographic data to individual files, numbered to correspond to the index number of the citing story, and then transferred to the spreadsheet. These data could then be analysed by a series of macros (all written by Philip Roe of Evaluametrics Ltd) to show the fractional counts of countries involved in authorship, the research level (from clinical to basic), the cancer anatomical site, and the type of research involved (*e.g*., epidemiology, surgery) (Pallari & Lewison, 2019).

Results

At this stage we have detailed results only for the few cancer research stories in *Global Times*, together with those in *New Scientist* for part of the six-year study period, and five previous years (2009-13). Because its objective is to cover mainly foreign news, the *Global Times* cited overseas newspapers frequently. Overall, 36% of the research that it cited was from the UK, 24% from the USA, and only 16% from China (including Hong Kong, and, curiously, Taiwan). Its stories were often focused on the relationship between lifestyles and cancer risks (48% of stories) and eye-catching medicines (24%). Screening and diagnosis were relatively rarely reported (8%). Much of the research mentioned was applicable to any type of cancer (44%).

There were more cancer research stories (n = 318) from the *New Scientist*. Research from the UK itself was over-cited (compared with its integer-count presence in all cancer research in the same years) by a factor of 4.5, and most European countries' research was also over-cited, by between 1.4 and 4 4 times. That of the USA was over-cited by a factor of 2.1, but research from East Asia was relatively ignored, with Chinese cancer research receiving only one quarter as many mentions (13 instead of 53) as would have been expected.

Stories within *New Scientist* over-emphasised breast cancer, with 33 of them mentioning it (10%), whereas it accounts for less than half the burden measured in DALYs than lung cancer (WHO, 2015). However, it is the most researched cancer site, both in the UK and the majority of other industrial countries (Begum et al., 2018). Other cancer sites prominently featured were the prostate (22) and the brain and central nervous system (21). The stories also concentrated on genetics, with 54 (17%) the most popular research domain, followed by drug treatment (40) and epidemiology (36). There were very few stories on the two other main cancer treatments, namely surgery (6) and radiotherapy (5), reinforcing the public view that drugs, and not these two latter treatments, are the way to defeat cancer.

Discussion and conclusions

China and the UK are very different countries, but they have a common interest in reducing the toll of cancer on their citizens. Since about half of the causes of cancer are to some extent under their control, it is important to see how news of the research that should underpin both their decisions and governmental regulatory actions is being presented to the public. This is an important component of the relationship between research and its public acceptance, which is one of the criteria for research evaluation in the forthcoming UK Research Excellence Framework for 2021 (https://www.ref.ac.uk

We have selected a wide range of newspapers and a popular science magazine in each country to study. The major part of the work remaining will be to collect all the details of the media stories and then of the papers that they cite in a spreadsheet; the analysis can be done very quickly with existing software. Once the results are clear we plan to conduct interviews with some of the key actors, such as science journalists, cancer researchers, and the funders of this research, in order to learn more about how research articles are selected for coverage, any constraints on this, how commentators on the new results are chosen, and perceptions of the impact the stories may have on the various groups of readers.

A limitation of the study is the relatively small number of newspapers and magazines that can be covered, and in the UK the lack of regional diversity. However, this is almost inevitable because of the decline of regional and local media. In China, the strict censorship imposed by the CCP makes this less of an issue, although we might still find some variation in the content of the stories in provinces where the cancer burden, or its treatment, differ markedly from that in the major cities of the east of the country.

References

Begum M, Roe P, Webber R & Lewison G. (2017) UK ethnic minority cancer researchers: their origins, destinations and sex. *Proceedings of the 16th Meeting of the International Society of Scientometrics and Informetrics*, Wuhan, China: 568-579.

Chen, W., Xia, C., Zheng, R., *et al.* (2019). Disparities by province, age, and sex in site-specific cancer burden attributable to 23 potentially modifiable risk factors in China: a comparative risk assessment. *The Lancet Global health,* 7**,** e257-e269.

Global Times, G. (2019). About Global Times [Online]. Available at: http://hd.globaltimes.cn/html/abouthq/ [Accessed January 2021].

Lee, C.-C. 2010. Bound to rise: Chinese media discourses on the new global order. *Reorienting global communication: Indian and Chinese media beyond borders***,** 260-283.

Lewison G, Tootell S. Roe P & Sullivan R. (2008) How do the media report cancer research? A study of the UK's BBC website. *British Journal of Cancer*, 99: 569-576

Pallari E & Lewison G. (2019) How Biomedical Research Can Inform Both Clinicians and the General Public. *Springer Handbook of Science and Technology Indicators* (W Glänzel et al., editors), 581-607. DOI: 10.1007/978-3-030-02511-322

Pallari E, Sultana A, Williams C & Lewison G (2020) An assessment of the coverage of non-communicable disease research reported in British and Irish newspapers, 2002-13 *Cogent Medicine*, 7:1757566

People's Daily, P. S. (2017). About Us [Online]. Available at: http://en.people.cn/90827/90828/ [Accessed February 2021].

Phillips DP, Kanter EJ, Bednarczyk B & Tastad PL. (1991) Importance of the lay press in the transmission of medical knowledge to the scientific community. *New England Journal of Medicine,* 325: 1180-1183.

Wang, Y. 2018. The Former Doctor Fighting China's Health Rumor Epidemic [Online]. Available: http://www.sixthtone.com/news/1001612/the-former-doctor-fighting-chinas-health-rumor-epidemic [Accessed February 2021].

World Health Organization (WHO) (2015) Available from: http://www.who.int/healthinfo/global_burden_disease/estimates/en/index2.html

Visualising plural mappings of science for Sustainable Development Goals (SDGs)

Ismael Rafols[1], Ed Noyons[2], Hugo Confraria[3] and Tommaso Ciarli[4]

[1] *i.rafols@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University (The Netherlands) &
Science Policy Research Unit (SPRU), University of Sussex, Brighton (England)

[2] *noyons@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University (The Netherlands)

[3] *h.confraria@sussex.ac.uk*
Science Policy Research Unit (SPRU), University of Sussex, Brighton (England)

[4] *t.ciarli@sussex.ac.uk*
Science Policy Research Unit (SPRU), University of Sussex, Brighton (England)

Abstract

Analysts are rapidly developing methods to map publications to SDGs in the face of policy demands. However, as reported by Armitage et al. (2020), a high degree of inconsistency is found when comparing the bibliometric corpora obtained with different approaches. These inconsistencies are not due to minor technical issues, but instead they represent different interpretations of SDGs. Given the variety of understandings regarding the relationship between research and SDGs, we propose that bibliometrics analysts should not assume that there is one single, preferred or consensus way of mapping SDGs to publications. We propose instead that, since different stakeholders have contrasting views about the relationships between science and SDGs, the contribution of bibliometrics should be to provide a plural landscape for stakeholders to explore their own views. We describe here the beta-version of an interactive platform that allows stakeholders to scrutinise in a global map of science the clusters potentially related to SDGs.

Introduction

The shift in S&T policy from a focus on research quality towards societal impact has led to a demand for new S&T indicators that capture the contributions of research to society (Wilsdon et al., 2017), in particular those aligned with SDGs. The use of the new 'impact' indicators would help monitoring if (and which) research organisations are aligning their research towards certain SDGs.

Responding to these demands data providers, consultancies and university analysts are rapidly developing methods to map projects or publications related to specific SDGs. These 'mappings' do not analyse the actual impact of research, but hope to capture instead if research is directed or related towards problems or technologies that can contribute to improving sustainability.

Yet this quick surge of news methods raises questions about the robustness of the mappings and indicators produced, and about the effects of using questionable indicators in policy making. The misuse of indicators in evaluation has been one of the key debates in science policy this last decade, as highlighted by initiatives such as the San Francisco Declaration on Research Assessment (2013) and the Leiden Manifesto (2015).

In a new study that aims to map publications to SDGs, we have found a high degree of inconsistency using different approaches, as recently reported by Armitage et al., (2020). In this contribution, we propose that these inconsistencies are not due to minor technical issues, but instead they represent different interpretations of SDGs. In other words, there is no a single objective 'truth' about which research is relevant for reaching SDGs. Since different stakeholders have contrasting and often conflicting views about the relationships between science and SDGs, the contribution of bibliometrics should be to provide a plural landscape for each stakeholder to explore which areas are 'really' related to SDGs according to his views

(Rafols and Stirling, 2020). We describe here the beta-version of an interactive platform that allows stakeholders to explore in a global map of science the relationship between publications and SDGs.

Recent efforts for mapping research to SDGs

The first public analysis of SDG impact, released in 2020 by the Times Higher Education (THE, 2019), should be a motive for concern. For almost two decades, the THE has offered a controversial ranking of universities according to 'excellence'. The THE has now produced a new ranking of universities according to an indicator that sums dimensions of unclear relevance. For example, the indicator of the impact on health (SDG3) of a university depends on the one hand on its relative specialisation on health (as captured, e.g. by the proportion of papers related to health, 10% of total weight), and on the other hand on the proportion of health graduates (34.6%). The score is also based on (self-reported) university policies such as care provided by the university, e.g. free sexual health services for students (8.6%) or community access to sports (4%). Such an heterogenous and arbitrary composite indicator is likely to cause more confusion than clarity and it is potentially harmful as it mystifies university policies for the SDGs.

One may thus believe that it is better to stick to objective measures such as 'relative specialisation on health' as captured by the proportion of papers related to health. In the THE ranking, this bibliometric data is supported by an Elsevier analysis of the publications that are related to the SDGs – which might seem more reliable than those based on data self-reported by universities (Jayabalasingham et al., 2019).

However, mapping publications to the SDGs is not straightforward. Bibliometricians are aware that while some consensus can be soon reached regarding the delineation of a traditional scientific fields, fields defined by policy issues such as environmental research are ambiguous. Following policy demands, there is currently a profusion of initiatives aiming at mapping publications to SDGs. The approaches developed by Bergen University (Armitage et al., 2020), Elsevier (Jayabalasingham, 2019), the Aurora Network, SIRIS Academic or the STRINGS project are based on searching for strings of keywords, in particular keywords found in the UN SDGs targets or other relevant policy documents. These searches are then enriched differently in each case. The hypothesis is that publications or projects containing these keywords are those best aligned with the UN SDG discourse. The question is then which keywords should be included, and which not. For example, why in some lists zika virus is included in the list of health SDG3, but not the closely related dengue virus, with a much higher disease burden?

An alternative approach being developed at NESTA and Dimensions (Wastl, 2020) uses policy documents and keywords to train machine learning algorithms in order to identify articles related to the SDGs instead of creating a list of keywords to search the articles. The alleged advantage of this algorithm is that machine learning is assumed to be more accurate than search strings. The downside of this approach is that is it a black box regarding the preferences (or biases) of the machine learning algorithms – and this is serious handicap for transparency.

In terms of consistency across approaches, an article recently published by a team at Bergen University sounded the alarm by showing that slightly different methods may produce extremely different results (Armitage et al., 2020). When comparing the papers related to SDGs retrieved with their own analysis with those by Elsevier, they found that there is astonishingly little overlap – in most SDGs only around 25%-35% as illustrated in Figure 1. The differences also affect the rankings of countries' contributions to the SDGs. The Bergen team concluded that 'currently available SDG rankings and tools should be used with caution at their current stage of development.' We have conducted exploratory comparisons between the Bergen, SIRIS and STRINGS approach and can confirm major differences.

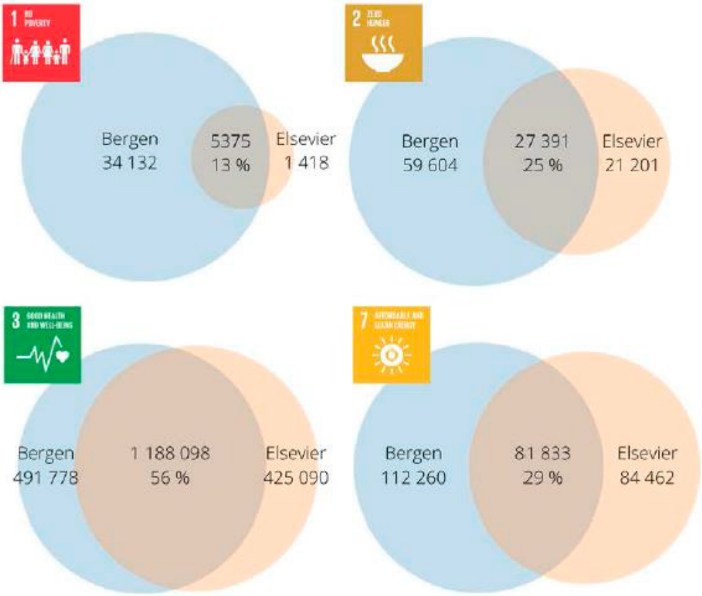


Figure 1. Comparison between Bergen and Elsevier approaches to mapping SDG-related publications. Based on WoS Core collection, 2015-2018. Source: Armitage et al. (2020).

Conceptual framework: a multiplicity of possible mappings of SDGs

However, from a conceptual point of view, perhaps we should not be surprised that different methods yield so different results. The SDGs refer to policy objectives in multiple dimensions – ending poverty, improving health, achieving gender equality, preserving the natural environment, etcetera. Innovation studies have shown that the social contributions of research are often unexpected and highly dependent on the local social contexts in which knowledges are created and used.

Nevertheless, most research is funded according to the expectations of the type of societal benefits that it may generate – and thus one can try to map these expectations or promises according to the language used in the (titles and abstracts of) projects and articles. However, the expected social contributions are often not made explicit in these technical documents because the experts reading them are assumed to know the potential applications.

As a consequence, the process of mapping projects or articles to the SDGs is ineluctably carried out through an interpretative process that 'translates' (or attempts to link) scientific discourse into potential societal outcomes. In other words, the analyst has to make assumptions regarding which terms are uniquely related to SDGs. Of course, such translation is dependent on the analysts' understandings of science and the SDGs. There is consensus on some of these understandings. For example, most analysts would agree that research on malaria is important for achieving global health. However, other translations are highly contested: should nuclear energy research be seen as a contribution to clean and affordable energy? Should all educational research be counted as contribution to the SDG on 'quality education'?

Furthermore, in a number of SDGs such as zero hunger (SDG 2) or reduced inequalities (SDG 10), there are stark disagreements on what the benefits of potential contributions: some stakeholders believe that GM crops (or organic farming) will help, while others believe that they will make the situation worse. Moreover, there is relatively little research explicitly mentioning issues such as gender or inequality in comparison to research whose innovation outcomes may affect these issues.

Another challenge is that many bibliometric databases used for analysis are not comprehensive, having a much larger coverage of traditional fields and rich countries (see Chapter 5 in Chavarro, 2017). This is particularly problematic, given that some research agendas relevant to developing countries are likely to be highly underrepresented (Rafols et al., 2015).

In summary, there is lack of consensus and many ambiguities on how publications relate to SDGs, and in these cases, the mappings will depend on the particular interpretation of the SDGs adopted in the methods for searching publications.

Given these ambiguities, we have developed an interactive mapping tool that aims to help stakeholders pick and choose which research lines they consider relevant in their context for each SDG. The map shows publication clusters of potentially relevant publication clusters for each SDGs with some interaction for users to understand the research contents of each cluster. This approach is introduced in the following sections.

Data and methods

We developed a search string extracting terms from policy documents explicitly mentioning SDGs. These search-strings were applied to retrieve publications (articles and reviews) for the period 2015-2019 from the CWTS in-house version of Web of Science (WoS) after an iterative process of refinement to avoid false negatives. This process was informed by comparisons with searches carried out by other initiatives, in particular SIRIS and Bergen. Details of this retrieval methodology are presented in the 2021 ISSI proceeding by Confraria et al.

On the other hand, we rely on an article-level classification based on direct-citation clustering of about 4,000 categories. These clusters are positioned in a science map based on citations across clusters according to a *VOSviewer* layout. This information is imported on a *Tableau* visualisation interface. The *Tableau* interface shows the clusters retrieved for each SDG with information on the cluster and the publications related to SDGs, as detailed below.

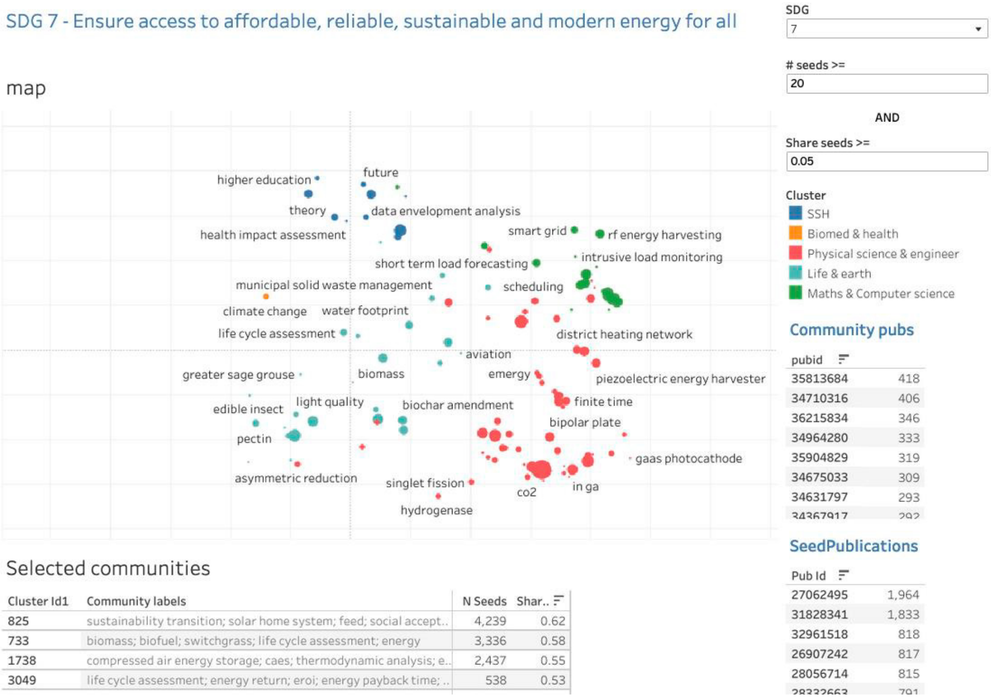
Results: a visualisation interface to explore plural mappings of SDGs

In this proceeding we introduce the beta-version of a visualisation interface aimed at facilitating the exploration of publication clusters related to SDG. The interface allows to choose one SDG (top right). Then the user can set the thresholds for considering if a cluster is related to this particular SDG: the minimum number of publications retrieved by the search string of this SDG (seed publications), and the minimum share of seeds in a cluster.

The clusters with figures above these thresholds are displayed in a science map according to their position in the global map of science. The colours of the clusters represent broad disciplines. Clusters in *Social science and humanities* (blue) are in the top left, *Biomedicine and health* (yellow) are in the left, *Physical sciences and engineering* (red) are in the bottom right, *Maths and computer sciences* (green) are in the right, and *Life and earth sciences* (light-blue) are in the centre-left.

The contents of specific clusters can be explored by placing the pointer of the mouse over the cluster of interest. For each cluster, you can see the 5 most relevant labels (where the labels are terms with high frequency and high cluster specificity) and the 5 most frequent journals. By clicking on the node or by clicking on the corresponding row in the list of 'Selected communities', the right column 'Community Pubs' allows to click into specific publications, which are linked via DOI to website of the publication. In this way, it is possible to read titles and abstracts and better understand the content of clusters.

Let us take for example, the cluster labelled as 'greater sage grouse' (centre-left) in SDG 7 (energy). This may seem a false positive since a grouse is a type of bird, rather than energy. However, the other labels of the cluster are wind farm and turbine. By clicking at the community publications, we then see that the article contents is related to bird collisions with wind turbines and power lines. Interestingly this cluster is about energy, but it shows a trade-off between SDG7 (Energy) and SDG15 (Life on Land).



SDG 7 - Ensure access to affordable, reliable, sustainable and modern energy for all

map

SDG
7 ▾

seeds >=
20

AND

Share seeds >=
0.05

Cluster
- 🟦 SSH
- 🟧 Biomed & health
- 🟥 Physical science & engineer
- 🟩 Life & earth
- 🟩 Maths & Computer science

Selected communities

Cluster Id1	Community labels	N Seeds	Shar..
825	sustainability transition; solar home system; feed; social accept..	4,239	0.62
733	biomass; biofuel; switchgrass; life cycle assessment; energy	3,336	0.58
1738	compressed air energy storage; caes; thermodynamic analysis; e..	2,437	0.55
3049	life cycle assessment; energy return; eroi; energy payback time; ..	538	0.53

Community pubs

pubid	
35813684	418
34710316	406
36215834	346
34964280	333
35904829	319
34675033	309
34631797	293
34367917	292

SeedPublications

Pub Id	
27062495	1,964
31828341	1,833
32961518	818
26907242	817
28056714	815
28332663	701

Figure 1. Interactive visualisation interface of the STRINGS project for SDG7 (Energy). It shows clusters with at least 20 and 5% of its publications having terms related to SDG7. Cluster labels are shown at the bottom left. Instances of publications of a cluster and those retrieve (seeds) shown in the right columns. Available at: https://public.tableau.com/profile/ed.noyons#!/vizhome/UKStringsSDGtocommunities/Dashboard1

The map illustrates the many different types of research trajectories that may be related to sustainable energy, from higher education (top left), to smart grid management (top right) to thin solid films for photovoltaics made out of copper indium gallium selenide solar cells (InGa, bottom right), to diesel engines (centre-right) and alternative fuels in aviation (centre). This large variety of research trajectories shows that science can contribute to sustainable energy in very different (sometimes conflicting) innovation pathways.

However, stakeholders in specific contexts may think that some research trajectories are not as relevant for them as other trajectories. For example, developing countries might object to considering that cancer research as a priority for SDGs, given the relative under-investment in infectious diseases – the latter issues having a much higher burden for them. Therefore, the proposal of an interactive map is aimed at facilitating that different stakeholders find and choose those research trajectories that they consider y relevant for SDGs. Instead of having one single delineation of each SDG, the use of the interactive visualisation by stakeholders makes it possible to develop a plural and conditional mapping between SDGs and publications (Rafols and Stirling, 2020).

We can thus think of research over a given SDG in terms of a variable portfolio of research options as shown by clusters, with different stakeholders having different perspectives, values and needs regarding the balance between options, and the options that should be prioritized (Wallace and Rafols, 2015; Ciarli and Rafols, 2019). A given SDG will not be achieved by having more research on this SDG – but by having a balance of research options that is conducive to achieving the SDGs. In contrast, we think that having only rankings of universities according to their SDG is not a meaningful source of information for policy decisions.

Conclusion

In summary, given the variety of understandings regarding the relationship between research and SDGs, we propose that bibliometrics analysts should not assume that there is one single, preferred or consensus way of mapping SDGs to publications. Such assumption is likely to reproduce hegemonic perspectives on SDGs – often given by rich countries at influential institutional actors (Stirling, 2019). Instead, we propose a mapping tool which suggests publication clusters potentially related to a given SDG, thus allowing plural interpretations.

Acknowledgments

This work is part of the STRINGS (STeering Research INnovation for Global GoalS). We are indebted to other members of the project for fruitful discussions. This proceeding is partly based on a blog post in the website of STRINGS at http://strings.org.uk/blog/.

References

Armitage, C. S., Lorenz, M., & Mikki, S. (2020). Mapping scholarly publications related to the Sustainable Development Goals: Do independent bibliometric approaches get the same results?. *Quantitative Science Studies*, *1*(3), 1092-1108.

Chavarro, Diego (2017) *Universalism and particularism: explaining the emergence and growth of regional journal indexing systems.* Doctoral thesis (PhD), University of Sussex. Retrieved on July 11 July 2020 from: http://sro.sussex.ac.uk/id/eprint/66409/

Ciarli, T., & Ràfols, I. (2019). The relation between research priorities and societal demands: The case of rice. *Research Policy*, *48*(4), 949-967.

DORA (2013) San Francisco Declaration on Research Assessment. Retrieved January 24, 2021 from: https://sfdora.org/

Jayabalasingham, B., Boverhof, R., Agnew, K., Klein, L. (2019) *Identifying research supporting the United Nations Sustainable Development Goals.* Elsevier documentation. DOI: 10.17632/87txkw7khs.1

Rafols, I., Ciarli, T., & Chavarro, D. (2015) Under-reporting research relevant to local needs in the global south. Database biases in the representation of knowledge on rice. Retrieved February 9, 2021 from: https://osf.io/preprints/socarxiv/3kf9d/

Rafols, I., & Stirling, A. (2020). Designing indicators for opening up evaluation. Insights from research assessment. Retrieved February 9, 2021 from: https://osf.io/preprints/socarxiv/h2fxp/

Stirling, A. (2019). How deep is incumbency? A 'configuring fields' approach to redistributing and reorienting power in socio-material change. *Energy Research & Social Science*, *58*, 101239.

THE (2020) Impact rankings. Retrieved February 25, 2021 from: https://www.timeshighereducation.com/impactrankings

Wallace, M. L., & Rafols, I. (2015). Research portfolio analysis in science policy: moving from financial returns to societal benefits. *Minerva*, *53*(2), 89-115.

Wastl, Juergen; Porter, Simon; Draux, Hélène; Fane, Briony; Hook, Daniel (2020): *Contextualizing Sustainable Development Research*. Digital Science. Report. https://doi.org/10.6084/m9.figshare.12200081.v2

Wilsdon, J. R., Bar-Ilan, J., Frodeman, R., Lex, E., Peters, I., & Wouters, P. (2017). Next-generation metrics: Responsible metrics and evaluation for open science. European Commission, Brussels. https://ec.europa.eu/research/openscience/pdf/report.pdf

Protégé-advisor gender pairings in academic survival and productivity of German PhD graduates

Andreas Rehs[1]

[1]*rehs@incher.uni-kassel.de*
University of Kassel, International Center for Higher Education Research, Mönchebergstr. 17, D-34127 Kassel;
+49-561-804-1936

Abstract

Female underrepresentation is a severe problem in academia and challenges the principle of merit-based selection. Previous studies have found protégé-advisor gender pairings to be associated with scientific survival and early career productivity for young scientists. In this study, I add to these results and focus on the disentanglement of temporal patterns after completing a PhD. I link a dataset of German protégé-advisor pairs scraped from online dissertations with disambiguated publication data from the Web of Science. My analysis is based upon time-discrete survival regression and explores the duration until the final publication after completing a PhD. I show that for protégés of female advisors, the yearly risk to drop out of academia after completing a PhD is about 38% lower than for protégés of male advisors. This effect is not different between female and male protégés. Women generally have a higher yearly dropout rate after completing a PhD.

Introduction

Doctoral advisors are often the most influential persons at the beginning of an academic career. They transfer knowledge, attitudes, norms, and behaviour to their protégés and influence their academic socialization and success (Barnes & Austin, 2009). Several studies have addressed the various scientific and socioeconomic characteristics of the advisors and their protégés to point out what makes these relationships mutually successful. Gender pairing in protégé-advisor relationships repeatedly stands out in this regard. It has diverse effects on career attainment and publication output of protégés (Gaule & Piacentini, 2018; Hilmer & Hilmer, 2007; Pezzoni, Mairesse, Stephan, & Lane, 2016).

In this paper, I want to investigate the role of protégé and advisor gender in German PhD graduates' academic outcomes. Especially for pairings involving women, this issue is of high societal and scientific interest in Germany. As observed in other countries, women are underrepresented in advanced career stages in German academia (Larivière, Ni, Gingras, Cronin, & Sugimoto, 2013). Although they accounted in 2017 for 51.7% of graduate students, their share of PhD holders was 45.4%. Women account for only 25.6% of university professors in Germany (Statistisches Bundesamt, 2020). This departure of women from the academic workforce indicates a misallocation of talent (Acemoglu, 1995). The consequences of this departure imply decelerated scientific progress with negative spillovers to industry and the economy in general. Women may also be personally affected. If they are equally qualified with men to start and pursue an academic career, but at some point quit, their educational investment cannot be fully utilized (McGuinness, 2006).

Gaule and Piacentini (2018) argue that this underrepresentation of women in academia perpetuates itself through the lower availability of female advisors for female students. They argue that underrepresentation works through a productivity channel or a preference channel. In the productivity channel, students are less productive when collaborating with an advisor of the opposite gender. As productivity is generally the primary driver of academic career success, this leads to higher rate of dropout from academia for female PhD graduates who were advised by men. In the preference channel, the authors argue that working with an advisor of the opposite gender is less enjoyable and leads to lower career satisfaction and a higher chance of dropping out early. Gaule and Piacentini show that a PhD student's research productivity, and propensity to become faculty after graduating, are both related to the gender of the advisor.

In this paper, I build on Gaule and Piacentini's findings. In the first step, I test whether productivity during a PhD in German academia is also linked to protégé-advisor gender pairings. In the second step, I focus on the disentanglement of the temporal patterns related to career outcomes and advisor gender after completion of a PhD. From the temporal perspective, academic careers, and careers in general, are non-dichotomous processes. They include multiple decisions and promotions that differ in their duration and in their point of time. The investigation of fixed points in time, as done in Gaule and Piacentini (2018), does not exploit the temporal dimension to its full extent. In this sense, it is an open question of how long protégés in different gender pairings remain in academia and which dropout "risk" they assume after their PhD. My study tries to address this gap and therefore adds to the methodical and empirical literature on academic labour market exit of junior researchers.

The durations one remains in academia can be considered as survival times and allow to utilize related models such as Cox proportional hazard or complementary log-log regression (c-log-log). The c-log-log regression used in this paper estimates covariates' effect upon the time a specified event takes to happen and assumes time to be discrete (Tutz & Schmid, 2016). Therefore, I can investigate how an advisor's gender and other characteristics affect the time a PhD graduate remains in academia after finishing his or her PhD. A similar methodology was applied by Sabatier, Carrere, and Mangematin (2006) to investigate the time it takes for female and male postdocs to attain professorship.

In the subsequent section, I discuss previous findings on gender pairings in academic protégé-advisor relationships. In the Data and methods section, I present my three data sources: doctoral advisor information scraped from German online dissertations, the German National Library's catalog, and publication data from the Web of Science that I have previously disambiguated (Rehs, 2021a,b). In the Results section, I describe the specification of my econometric approach and use negative binomial regression to estimate the effect of advisor gender on PhD student productivity. While I found that women were less likely to publish during the PhD, being advised by women did not have any effect on publication productivity. The probability of academic exit from academia by gender and advisor gender, as measured by the last year of publication after completion of a PhD, is investigated with a c-log-log regression and represents my main finding. I find that PhD graduates who had female advisors are about 37% more likely per year to continue publishing after PhD completion; this effect is not different between male and female graduates. In line with the observable female underrepresentation in academia, I also find that women are about 38% more likely per year than men to exit from research after completing a PhD. I end this paper by discussing the results, showing limitations, and, finally, concluding.

Gender pairings and outcomes of doctoral advisory

The topic of gender in doctoral advisory belongs to the greater literature on protégé-advisor relationships. This literature is divided into business research (Feeney & Bozeman, 2008; Noe, 1988), undergraduate (Bettinger, Long, Ehrenberg, Jacob, & Murnane, 2005), and postgraduate protégé-advisor relationships. Central to all these literature strains is some success outcome, like establishing business networks (Feeney & Bozeman, 2008) or influencing women to major in scientific fields of female academic role models (Bettinger, Long, Ehrenberg, Jacob, & Murnane, 2005; Canaan & Mouganie, 2019). In summarizing the literature across all those subdomains, I find that there is no clear support for the hypothesis that female advisors positively affect the outcomes of their female protégés.

The relation of advisor and protégé gender in postgraduate outcomes has been addressed in several studies and is central to the debate of female underrepresentation in academia (Pezzoni et al., 2016). When discussing those studies, one must account for the various disciplinary, institutional, and regional backgrounds in which the studies were conducted. The German context – which is subject of this study – is special in many regards (Kehm, 2006). German universities produce one of the highest proportions worldwide of doctorates in relation to the population (OECD, 2019). About 50% of those doctorates are awarded in the field of medicine, where dissertations are very different from other disciplines (Konsortium Bundesbericht Wissenschaftlicher Nachwuchs, 2017).

The sorting process of advisors and their protégés is the natural starting point to investigate the effects of advisor gender and gender pairings. It has repeatedly been found that same-gender pairings are clearly overrepresented (AlShebli, Makovi, & Rahwan, 2020; Gaule & Piacentini, 2018; Pezzoni et al., 2016). The causal mechanisms for this overrepresentation, however, remain unclear. The qualitative study from Gray and Goregaokar (2010) on executive coaching suggests that women prefer women because they act as a role model for business success. This can lead to empirical problems as the matching process may not be random. Azoulay, Liu, and Stuart (2017) especially point towards this issue of "partially deliberate" social matching. Some social factors, like the rationales for forming the advisor-protegee relationship (e.g., same gender role models) are endogenous to the matching process. However, other factors are incidental and are therefore conditionally exogenous to the matching process. Azoulay, Liu, and Stuart (2017) indicate that geography and thematic focus are the main drivers of matching between advisors and protégés. For my study this implies that self-selection bias related to same gender can be present in the data but is probably not the main driver of matching.

The international literature comes to different conclusions on the effects of postgraduate advisor gender and protégé-advisor pairings. Starting with productivity outcomes, Pezzoni et al. (2016) find for protégé-advisor pairs of the prestigious California Institute of Technology that students working with female advisors publish 7.7% more articles per year while earning their PhD than those working with a male advisor. Using male students with male advisors as the reference group, Pezzoni et al. show that gender pairing matters in this regard. They find that male students working with female advisors publish 10% more articles per year than the reference group, and female students working with male advisors publish 8.5% less. They find no difference between women advised by women and the reference group of men advised by men. The results are robust for using the journal impact factor as a proxy for the quality of articles. Gaule and Piacentini (2018) study the productivity of PhD students in US chemistry programs. In opposition to Pezzoni et al., they find students with an advisor of the same gender tend to be more productive during a PhD program. Women profit more strongly from same-gender advisors than men do.

The recent study from AlShebli, Makovi, and Rahwan (2020) on protégé-advisor gender pairings is controversial (Retraction Watch, 2020). They use a different notion of advisor, considering all senior coauthors on a paper as the focal PhD student's informal advisors. Using 3 million protégé-advisor pairs, they find same-gender pairings do not benefit the impact of papers (measured by number of times cited by other papers) written after completing a PhD. Their findings suggest that female protégés who remain in academia profit more strongly when mentored by male advisors rather than equally impactful female advisors. Their findings also suggest that advisors more strongly benefit in future impact when working with male protégés rather than comparable female protégés. This especially holds if the mentor is female. The paper from AlShebli, Makovi, and Rahwan has been criticized because advisor quality, as measured by previous papers, is an inadequate matching variable. The advisor quality is central in finding comparable female-male, female-female, male-male, and male-female pairs that are later used for analysis. Since there is empirical evidence for gender citation bias, matching via this citation-based criterion may also be biased.

Hilmer and Hilmer (2007) and Neumark and Gardecki (1998) investigate protégé-advisor gender pairings in economics and consider activity-based success measures. When examining the first jobs of new PhD graduates, Hilmer and Hilmer (2007) find that female graduates who had male advisors are significantly more likely to accept research-oriented first jobs than male graduates who had male advisors. Neumark and Gardecki (1998) focus on time spent in and completion of graduate school. They find limited empirical evidence for the positive impact of female advisors on the probability that female students finish graduate school. However, female advisors are associated with female students spending less time in graduate school. Gaule and Piacentini (2018) address the likelihood of PhD students becoming university faculty based on different gender pairings. They find female students working with female advisors are considerably more likely to become faculty; for male-male pairings they do not find an effect. In summary, the empirical evidence on the effects of gender pairings and advisor gender is ambiguous, scattered, and may depend strongly on the data, context, and operationalization of outcomes. I therefore abstain from forming a hypothesis about the protégé-advisor gender pairing effects in German academia.

Data and methods

My data rest upon two pillars: disambiguated Web of Science publication data, and PhD advisor info scraped from online dissertations. A schematic of my databases and their relations is depicted in Figure 1. I use the German National Library's 2015 electronic catalog and university library servers to build my online dissertations database. The German National Library has the legal mission to collect and archive all printed publications issued in Germany and works written in German or relating to Germany. German PhD graduates are therefore required to supply a copy of their dissertation to the German National Library. The German National Library's electronic catalog features information on their authors, the university name, the year of publication, subject, and, if available, a link to an online dissertation.

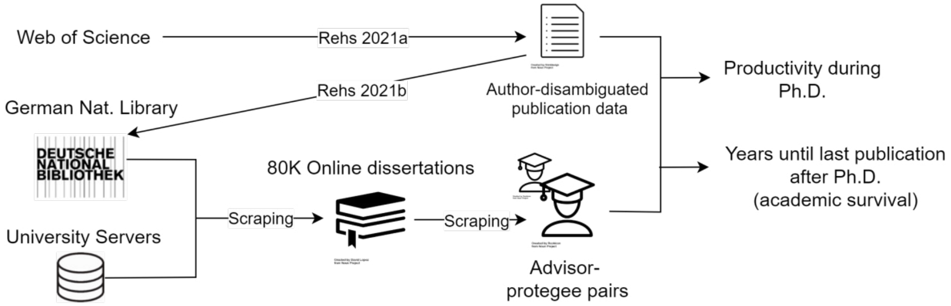

Web of Science
Rehs 2021a
Rehs 2021b
German Nat. Library
DEUTSCHE NATIONAL BIBLIOTHEK
Author-disambiguated publication data
Productivity during Ph.D.
80K Online dissertations
Years until last publication after Ph.D. (academic survival)
Scraping
Scraping
University Servers
Advisor-protegee pairs

Figure 1. Database schematic

I use the provided online dissertation link and download the underlying PDF document. The download from the German National Library was successful in 40,000 cases. To further increase my online dissertation dataset, I repeat the same exercise for dissertations stored at 20 different university library servers, collecting 80,000 online dissertations. All scraping has been done in 2017. I match the 40,000 dissertations from university library servers back to the German National Library's catalog by the author name, year, and university name. I search dissertation front pages and acknowledgments for text patterns like "doctoral advisor" and its German variants in the next step. These patterns indicate the subsequent occurrence of advisor info. A similar approach was used by Fuchs and Rehs (2020) to scrape birthplaces from the same dataset of online dissertations. I find 13,315 protégé-advisor pairs where the found advisor has a unique name in the German National Library catalog. The restriction to unique advisor names ensures correct protégé-advisor pairs.

To retrieve publication data of the protégés and their advisors, I use the database from Rehs (2021b). Rehs (2021b) builds on Rehs (2021a), which develops a machine learning approach to disambiguate author names in the Web of Science. Rehs (2021b) uses this algorithm to establish a data base of about 11 million author-name disambiguated publications and links them to 50,000 out of one million dissertation authors stored in the German National Library's catalog. Using the publication profiles of German dissertation authors, I investigate how long they continue to publish in Web of Science journals after completing their PhD and how this duration is related to their gender and the gender of their advisor.

Productivity during the PhD and other related indicators are also calculated from this dataset. For a random set of 100 persons, I check whether the year of the final publication after completing a PhD corresponds to the actual year of dropout from academia. I retrieve the dropout year from online research on Xing, LinkedIn, university homepages, and other websites. The year of the final publication correlates with the actual dropout by 0.7 and validates the year of final publication as a dropout proxy.

Table 1 reports summary statistics for my final datasets. In line with Gaule and Piacentini (2018), I find that women disproportionally often advise women. Advisors also differ substantially in their numbers and their characteristics by gender, although advisors have, on average, the same academic age (see advisor characteristics: *Dissertation year*). My earliest dissertation from protégés is from 2001; the mean dissertation year is 2005. The availability of online dissertations is the driver of this bias towards younger doctoral cohorts in my dataset. The high popularity of online dissertations may explain the sharp increase in the coverage rate of advisor info in my dataset. The difference between the PhD graduate's dissertation year and the advisor's dissertation year (see advisor characteristics: *Difference to diss. year of protégé*) is a measure for the advisor's academic age. Female and male protégés have no differences in this regard.

Figure 2 shows the temporal distribution of the mean number of (accumulated) publications of a protégé by gender pairing. The descriptive patterns suggest that productivity differences are established early, before completion of a PhD. Male protégés advised by men have the highest mean productivity and women advised by women, the lowest. After $t = 0$ the publication number averages may include survivor bias and can therefore not be interpreted in a meaningful way.

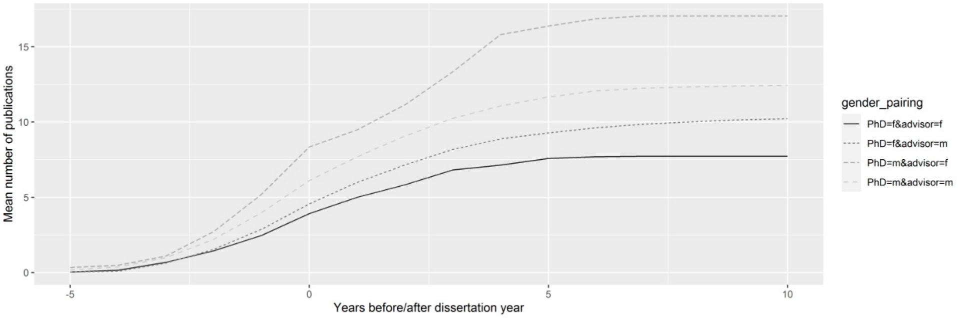


Figure 2. Mean number of cumulated publications before and after dissertation by gender pairing

Years till last pub. after diss. is my main outcome with respect to academic survival. I can observe that PhD graduates advised by men survive on average half a year longer than those by female advisors. In further differentiating the effect of female advisors, I find that their female students produce their final publication half a year earlier than male students. I observe no difference between male and female PhD graduates with male advisors in the mean number of years until the final publication.

Table 1. Descriptive statistics

Protégé characteristics												
		Full sample				Protégé = male				Protégé = female		
	N	min	mean	max	n	min	mean	max	n	min	mean	max
Gender advisor = m.	873				499				374			
Gender advisor = f.	89				21				68			
Years in academia after dissertation	962	0	2,5	15	520	0	2,5	15	442	0	2,4	15
Years till last pub. after diss. & advisor = m.	873	0	2,5	15	499	0	2,5	15	374	0	2,4	15
Years till last pub. after diss. & advisor = f.	89	0	2,1	7	21	0	2,5	7	68	0	1,9	7
Dissertation year		2001	2011	2015		2001	2011	2015		2001	2011	2015
Number of papers till 2017		1	11.5	118		1	12,6	72		1	10,2	118
Sum of papers at dissertation year		0	5,4	66		0	6,2	66		0	4,6	30
Sum of papers at diss. year & advisor = m.	873	0	5,4	66	499	0	6,1	66	374	0	4,6	30
Sum of papers at diss. year & advisor = f.	89	0	5,0	39	21	0	8,33	39	68	0	3,9	14
Number of citations		0	150	555		0	154	555		0	108	1,233
Sum of citations at dissertation year		0	6,3	78		0	6,5	78		0	5,2	63
Advisor characteristics												
		Full sample				Advisor = male				Advisor = female		
	N	min	mean	max	n	min	mean	max	n	min	mean	max
Dissertation year	962	1913	1990	2015		1913	1990	2015		1966	1991	2007
Difference to diss. year of protégé	962	0	20,5	41		-4	21,4	99		6	20,4	41
Number of papers	59	2	59,1	208		2	63,4	189		6	52,9	208

In my econometric setup, I will first investigate a student's productivity during the PhD program to address whether early productivity differences by gender and gender of the advisor exist. The number of papers written until the year after the PhD is my dependent variable. The main variables of interest are the gender of the protégé and the gender of the advisor. Since the outcome variable is a count, I use a negative binomial regression to estimate my regression and control for discipline and year effects using dummy variables.

The next step is modeling academic survival after completing a PhD. I use the year of the final publication after PhD completion as a proxy for academic survival. This outcome's operationalization is restricted in my dataset by the period of graduation cohorts from 1995 to 2015. Older cohorts can therefore be active longer than younger cohorts. My solution is to censor persons who were still active in 2017. A PhD graduate from 2014, still publishing at the cutoff of my database in 2017, is treated as right-censored after three years. In the following description of my economic approach, I will orientate on the methodology of Tutz and Schmid (2016) and van de Schoot (2020).

In my time discrete survival models the risk of event builds on the hazard h_{it}. The hazard is the conditional probability that a researcher i will exit from academia in the time period t, given the researcher did not exit earlier. The hazard function can be stated as follows:

$$h_{it} = P(T_i = t \mid T_i \geq t) \tag{2}$$

In (2), T is a discrete random variable. The equation represents the probability that the exit of a given researcher will occur in the current time period t under the constraint that it will occur now or sometime in the future. Now, the hazard in time period t can be estimated as follows:

$$\hat{h}_t = \frac{Number\ of\ exits\ from\ research_t}{Number\ of\ researchers\ at\ risk\ to\ exit_t} \tag{3}$$

In order to build a regression framework, the hazard h_{it} now needs to be linked to a linear predictor η. The relationship of the hazard function and a linear predictor can be represented as:

$$\eta = g(h_{it}) = \gamma_{0t} + x_{it}\gamma \tag{4}$$

Here, g is a link function that links the linear predictor to the hazard. In my case, the linear predictor includes a set of covariates for researcher i in period t. These covariates are protégé gender, advisor gender, and year and discipline controls. To disentangle the effects of advisor gender, I also estimate interactions of advisor gender and protégé gender. In (4), γ_{0t} represents the time-variant intercept and shows the baseline hazard. In the next step, I use a c-log-log function to link the hazard to the linear predictor. Therefore, my model refers to a time discrete c-log-log regression. The hazard function changes and becomes:

$$h_{it} = 1 - \exp(-\exp(\eta)) \tag{5}$$

Results

Table 3 shows the regression results and average marginal effects for productivity as measured by the number of papers published during PhD study. I do not find an effect of advisor gender in any of my full sample models. However, the coefficients and the marginal effects of protégé gender arrive at statistical significance in the baseline, in the full model, and in the full model with advisor productivity. According to the full model's marginal effects, female PhD students write 1.6 papers less than male PhD students. When including advisor productivity, as done in the full model 5, this discrepancy disappears. The advisor productivity has a statistically significant impact on the number of papers at the dissertation year. An additional publication from the advisor leads to an increase of about 0.04 publications by their protégés. Since this result is based on 59 observations, including only nine female advisors, the robustness is questionable.

Table 3. Negative binomial regression and average marginal effects for productivity during PhD study

	Dependent variable: Number of papers during PhD									
	(1)		(2)		(3)		(4)		(5)	
PhD student characteristics	Coeff.	AME	Coeff.	AME	Coeff.	AME	Coeff.	AME	Coeff.	AME
Gender PhD student = female	- 0.327*** (0.075)	-1.767*** (0.419)	- 0.291*** (0.078)	- 1.786*** (0.421)	- 0.262*** (0.076)	- 1.417*** (0.419)	- 0.2288** (0.079)	- 1.436*** (0.421)	-0.4743 (0.327)	-3.476 (2.492)
Advisors gender = female	0.009 (0.129)	-0.049 (0.700)	0.311 (0.245)	0.728 (0.885)	-0.036 (0.127)	-0.198 (0.686)	0.268 (0.239)	0.526 (0.854)	-0.988* (0.418)	-7.247* (3.341)
Advisors gender = female PhD student = female*			-0.4646 (0.289)	-			-0.4399 (0.2831)	-		
Difference to dissertation year advisor					-0.006 (0.0037)	-0.032 (0.020)	0.0060 (0.003)	-0.033 (0.020)	-0.0439* (0.019)	-0.0322* (0.157)
PhD advisor characteristics										
Advisor productivity									0.0053* (0.002)	0.0396* (0.018)
Discipline dummies	NO		NO		YES		YES		YES	
Year dummies	NO		NO		YES		YES		YES	
Observations	961		961		961		961		59	

Note: **p* < 0.1; ***p* <0.05; ****p* <0.01

Figure 3 shows the Kaplan Meier survival curve for the four different protégé-advisor constellations and a table that shows the number of graduates at risk of leaving academia each year after PhD completion (t = 0). I observe a substantial decline in the number of persons at risk in the first two years. The dominant share of scientists, therefore, do not stay in academia after PhD. P = 0.21 indicates the log-rank test result and indicates that the time to the final publication is statistically not different between the four groups.

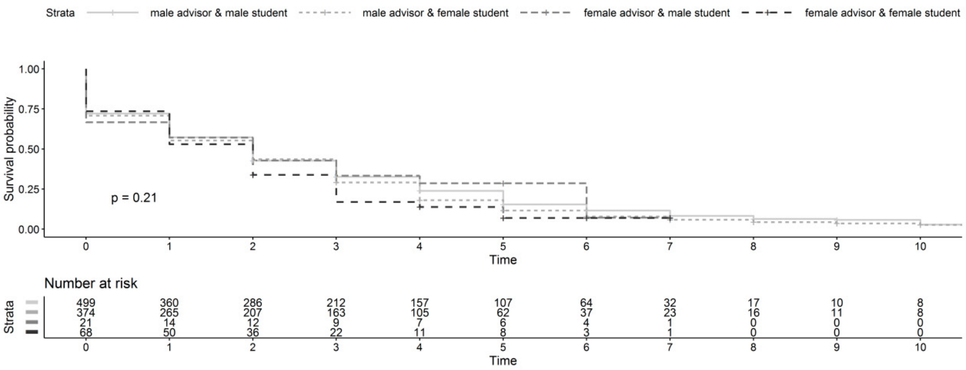


Figure 3. Kaplan Meier Curve and risk table for final publication after PhD completion

Table 2 shows the c-log-log regression results. The hazard for women to write their final paper is, according to Model 1, 1.37% higher than for men. A female advisor is generally beneficial and leads to a 38% lower hazard of dropout from academia. Model 3 includes year and discipline dummies; it shows that women's higher hazard remains robust when including these dummies. Models 2 and 4 display the interaction of advisor gender and protégé gender. I find no statistically significant effect for the interaction.

Table 2. C-log-log model – Yearly hazard of exit from research

	Dependent variable: Yearly hazard of exit from research							
	(1)		(2)		(3)		(4)	
	Coef.	Exp (Coef.)	Coeff	Exp (Coef.)	Coef.	Exp (Coef.)	Coef.	Exp (Coef.)
PhD student characteristics								
Gender PhD student=female	0.315*** (0.051)	1.37	0.335*** (0.054)	1.40	0.321*** (0.054)	1.38	0.343*** (0.056)	1.41
Difference to dissertation year advisor					-0.007** (0.002)	0.99	-0.007** (0.002)	0.99
Number of papers during PhD					-0.008** (0.003)	0.99	-0.008* (0.003)	0.99
PhD advisor characteristics								
Gender of advisor	-0.474*** (0.087)	0.62	-0.293* (0.156)	0.75	-0.481*** (0.090)	0.62	-0.264 (0.163)	0.77
Gender Interaction								
*Gender PhD student *advisor gender=female:*			-0.2935 (0.156)	0.77				
Intercept	1.680*** (0.063)	5.37	1.678*** (0.063)	5.36	2.006*** (0.104)	7.44	2.004*** (0.104)	7.42
Baseline risk	-0.196*** (0.009)	0.82	-0.197*** (0.009)	0.82	-0.203*** (0.010)	0.82	-0.204*** (0.010)	0.81
Discipline dummies	NO		NO		YES		YES	
Year dummies	NO		NO		YES		YES	
Observations	3324		3324		3321		3321	
Chi2	547.08		548.84		589.85		592.05	
Pseudo R2 Mc Fadden	0.19		0.19		0.21		0.21	
AIC	2334.1		2334.37		2310.34		2310.15	

Note: **p* < 0.1; ***p* <0.05; ****p* <0.01

Discussion and conclusion

In this paper, I have investigated the effect of advisor gender on protégé productivity during PhD study and on academic survival after PhD completion. In my investigation of productivity, I aimed to find if early productivity differentials are related to advisor gender. Using negative-binomial regression, I do not find any such relation. However, I find that female PhD students publish about 1.6 papers less than male students during their PhD study. My data cannot explain if this profound effect is related to other advisor and PhD student characteristics. For instance, it is still unknown if advisor productivity, quality, and protégé-advisor collaboration play a role. Also, I could not account for publication quality and number of co-authors. In future studies quality can be addressed by weighting the publications on the basis of the journal impact factor. My second outcome, the time between PhD completion and final publication, addresses academic survival after earning a PhD. The time until the final publication is one of the main contributions of my approach. Unlike in previous literature, which concentrates on examining outcomes at fixed points, I fully utilize the temporal dimension after PhD completion.

My application of time-discrete c-log-log regression delivers two main findings on advisor and protégé gender. First, I find that female advisors positively affect academic survival of PhD graduates, leading to a 38% lower dropout hazard. The causal mechanisms underlying this result remain unclear and are beyond the focus of my study. My results potentially suffer from self-selection bias. Therefore, the initial matching process between protegees and advisors may not be random. Descriptively, the disproportionately high number of women advised by women may indicate such bias. However, Azoulay et al. (2017) point out that spatial proximity and thematic focus are the main drivers for matching advisors and protégés. In further studies the potential self-selection bias should therefore be addressed by using more advanced econometric setups (e.g., Azoulay et al., 2017).

The female advisor effect is statistically not different between male and female PhD students. My results contrast with previous findings, such as those of Gaule and Piacentini (2018), who find that same-gender protégé-advisor pairings increase the likelihood of PhD graduates becoming university faculty. The reason for my lack of result may be a problem of low statistical power. PhD graduates advised by women make up only 89 observations in my dataset, and the effect just fails to reach the 10% statistical significance level. To obtain more observations from female advisors, one can improve the advisor scraping and linking and author name disambiguation strategy in the future. The underlying dataset of Rehs (2021b) only covers a fraction of German dissertation authors who have published.

An explanation of the low number of female PhD graduates and female advisors could also be attributed to name changes after marriage. My approach does not account for women who marry and change their family name after their doctorate. Since they are then no longer observable, they are considered to have had their final publication. This problem could be solved by bibliographic coupling. If I no longer observe publications from the focal female scientist, and if a previous coauthor of hers publishes together with a person of the same first name, but different last name, this may indicate that the focal scientist has changed her name. The positive female advisor effect may be conditional to unaddressed advisor characteristics. I did not control for team characteristics, informal advisors, graduate school characteristics, funding characteristics, socioeconomic background, and many others.

The second main result I find is that women are 37% more likely per year than men to exit from research after completing a PhD. In light of the female underrepresentation, this finding is unsurprising. Nevertheless, it adds to the literature by quantifying the female dropout risk for the first time. Concerning the name changes mentioned previously, this result needs further robustness analysis. There are other factors unaddressed in my study that lead to female dropout. In particular, the effect of private events and factors, such as childbirth and motherhood, lead to an omitted variable bias in my study. It is also unclear whether there are differences in the preference for careers in science or industry between men and women after completing their PhD, which could explain the female departure from academia.

Finally, my study is limited to young German scientists in disciplines where publication in Web of Science index journals is traditionally dominant. Therefore, I miss arts, humanities, and parts of social science where journal publications are not (or have not been) popular and which are only partly covered by the WOS (Mongeon & Paul-Hus, 2016). In these disciplines, women also make up higher shares of PhD graduates in Germany, potentially leading to different productivity and survival patterns.

I conclude that female underrepresentation and its relation to advisor gender in academia is a complex empirical phenomenon of high scientific and societal relevance. My study contributed to the literature by trying to disentangle the temporal and productivity dimensions before and after completion of a PhD in the German context. The survival operationalization offered a new perspective on female underrepresentation and may help to design policies that help to reduce female underrepresentation in academia.

References

Acemoglu, D. (1995). Reward structures and the allocation of talent. *European Economic Review*, *39*(1), 17–33.

AlShebli, B., Makovi, K., & Rahwan, T. (2020). The association between early career informal mentorship in academic collaborations and junior author performance. *Nature Communications*, *11*(1).

Azoulay, P., Liu, C. C., & Stuart, T. E. (2017). Social influence given (Partially) deliberate matching: Career imprints in the creation of academic entrepreneurs. *American Journal of Sociology*, *122*(4), 1223–1271.

Barnes, B. J., & Austin, A. E. (2009). The role of doctoral advisors: A look at advising from the advisor's perspective. *Innovative Higher Education*, *33*(5), 297–315.

Bettinger, E. P., Long, B. T., Ehrenberg, R., Jacob, B., & Murnane, R. (2005). Do faculty serve as role models? the impact of instructor gender on female students. *American Economic Review*, *95*(2), 152–157.

Canaan, S., & Mouganie, P. (2019). Female Science Advisors and the STEM Gender Gap. *SSRN Electronic Journal*. Elsevier BV. Retrieved December 11, 2020, from https://papers.ssrn.com/abstract=3396119

Feeney, M. K., & Bozeman, B. (2008). Mentoring and network ties. *Human Relations*, *61*(12), 1651–1676.

Fuchs, M., & Rehs, A. (2020). Career paths of PhD holders in eastern and western Germany Same qualification, same labor market outcomes? *IAB-Discussion Paper*, IAB-Discussion Paper, *2020*(1). Retrieved December 11, 2020, from http://doku.iab.de/discussionpapers/2020/dp0120.pdf

Gaule, P., & Piacentini, M. (2018). An advisor like me? Advisor gender and post-graduate careers in science. *Research Policy*, *47*(4), 805–813. Elsevier B.V.

Gray, D. E., & Goregaokar, H. (2010). Choosing an executive coach: The influence of gender on the coach-coachee matching process. *Management Learning*, *41*(5), 525–544. SAGE PublicationsSage UK: London, England. Retrieved December 12, 2020, from http://journals.sagepub.com/doi/10.1177/1350507610371608

Hilmer, C., & Hilmer, M. (2007). Women helping women, men helping women? Same-gender mentoring, initial job placements, and early career publishing success for economics PhDs. *American Economic Review, 97*(2), 422–426.

Kehm, B. M. (2006). Doctoral education in Europe and North America: a comparative analysis. *Wenner Gren International Series*, *83*, 67–78.

Konsortium Bundesbericht Wissenschaftlicher Nachwuchs. (2017). *Bundesbericht Wissenschaftlicher Nachwuchs 2017*. Bielefeld: W. Bertelsmann.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature*, *504*(7479), 211–213.

McGuinness, S. (2006). Overeducation in the labour market. *Journal of Economic Surveys*, *20*(3), 387–418.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228.

Neumark, D., & Gardecki, R. (1998). Women helping women? Role model and mentoring effects on female Ph.D. students in economics. *Journal of Human Resources*, *33*(1), 220–246. University of Wisconsin Press.

OECD. (2019). OECD work on careers of doctorate holders. Retrieved December 12, 2020, from https://www.oecd.org/innovation/inno/careers-of-doctorate-holders.htm

Pezzoni, M., Mairesse, J., Stephan, P., & Lane, J. (2016). Gender and the publication output of graduate students: A case study. *PLoS ONE*, *11*(1).

Rehs, A. (2021a). A supervised machine learning approach to author disambiguation in the Web of Science. *Journal of Informetrics* (forthcoming).

Rehs, A. (2021b). The scientific productivity of German PhD graduates: A machine learning based author name disambiguation and record linkage approach. To appear in *Proceedings of the 18th International Conference on Scientometrics and Informetrics.*

Retraction Watch. (2020). Nature Communications looking into paper on mentorship after strong negative reaction. Retrieved December 1, 2020, from https://retractionwatch.com/2020/11/19/nature-communications-looking-into-paper-on-mentorship-after-strong-negative-reaction/

Sabatier, M., Carrere, M., & Mangematin, V. (2006). Profiles of academic activities and careers: Does gender matter? An analysis based on french life scientist CVs. *Journal of Technology Transfer*, *31*(3), 311–324. Springer. Retrieved December 12, 2020, from https://link.springer.com/article/10.1007/s10961-006-7203-3

van de Schoot, R. (2020). Intro to Discrete-Time Survival Analysis in R. Retrieved February 18, 2021, from https://www.rensvandeschoot.com/tutorials/discrete-time-survival/

Statistisches Bundesamt. (2020). Frauenanteile nach akademischer Laufbahn. Retrieved October 29, 2020, from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Tabellen/frauenanteile-akademischelaufbahn.html

Tutz, G., & Schmid, M. (2016). *Modeling Discrete Time-to-Event Data*. Springer Series in Statistics. Cham: Springer International Publishing. Retrieved December 8, 2020, from http://link.springer.com/10.1007/978-3-319-28158-2

Reporting transparency in analysis of variance assumption testing

Maxime Sainte-Marie[1], Asger Dalsgaard Pedersen[2] and Philippe Mongeon[3]

[1] *msaintemarie@ps.au.dk*
Danish Center for Studies in Research and Research Policy, Aarhus University (Denmark)
Centre interuniversitaire de recherche sur la science et la technologie, Université du Québec à Montréal (Canada)

[2] *adp@ps.au.dk*
Danish Center for Studies in Research and Research Policy, Aarhus University (Denmark)

[3]*pmongeon@dal.ca*
Quantitative Science Studies Lab, School of Information Management, Dalhousie University (Canada)

Abstract

ANOVA models in general are valid only insofar as both homoscedasticity and multiple comparisons between groups are controlled for. While recent studies suggest that researchers in various disciplines seldom comply with these requirements, lack of rigor in methodological reporting makes a proper, large-scale, and transdisciplinary assessment of the situation impossible. Given this situation, the present paper attempts a large-scale and transdisciplinary assessment of statistical reporting transparency in ANOVA-related papers, using customized regular expressions on the full-text of articles published in Elsevier journals. Results show that, beyond important variations at multiple disciplinary levels, reporting transparency in the testing of ANOVA statistical assumptions is generally lacking, as a minority of ANOVA-related papers refer to either homoscedasticity or multiplicity correction. Also, articles from higher-quartile journals tend to report statistical assumptions more thoroughly than those from lower-quartile journals, regardless of discipline or subdiscipline.

Introduction

Alongside scientific misconduct, questionable research practices (QRPs) have the potential to undermine science, its reputation, and human knowledge in a considerable, extensive, and durable manner (John, 2012; Martinson et al., 2005, Fiedler and Schwarz, 2016; Sacco et al., 2018; Steneck, 2006). What makes QRPs such as post-hypotheses, conveniently-rounded numbers or selective literature reviews particularly insidious, resilient, and prone to escalation is that they can be done unknowingly, in good faith, and offer “considerable latitude for rationalization and self-deception” (John, 2012: 524): “QRPs can waste researchers’ time and stall scientific progress, as researchers fruitlessly pursue extensions of effects that are not real and hence cannot be replicated. More generally, the prevalence of QRPs (…) threatens research integrity by producing unrealistically elegant results that may be difficult to match without engaging in such practices oneself” (John, 2012: 531).

Amongst the different issues pertaining to QRPs, statistical misreporting certainly features amongst the most benign-looking, yet deceptive ones: despite the foundational and historical role of thorough reporting in science (Hardwicke, 2019), empirical research in certain fields has shown that reporting inconsistencies are increasing (Leggett, 2013), can have important consequences (Chalmers, 1990), and persist despite publication of guidelines (Clayson, 2019). Even when stemming from honest mistakes, selective or incomplete reporting of statistical protocols results in publications that “fail to capture all of the findings generated by the scientific enterprise” and thus provide “a skewed impression of the evidentiary landscape” (Hardwicke, 2019: 15). Perhaps more importantly, statistical misuses cannot be properly identified, assessed or dealt with if scientific narratives tell only part of the story.

This deadlock is perhaps nowhere more obvious than in the case of analysis of variance (ANOVA) procedures, which figure amongst the most popular Null Hypothesis Significance Testing (NHST) models in science. Regardless of the number of independent or dependent variables considered, from a parametric or nonparametric perspective, ANOVA procedures are only valid insofar as both homoscedasticity of variance and multiplicity of comparisons between groups are both controlled for. While studies presented in the following section show that both requirements are seldom met and that non-compliance in both regards can greatly impact research results, lack of rigor in methodological reporting makes a proper evaluation of the situation impossible. In light of this, the present research attempts a first large-scale and multidisciplinary assessment of homoscedasticity and multiplicity reporting in ANOVA-based research, using customized regular expressions on the full-text of articles published by Elsevier journals in 2015. In the next section, the different types of analysis of variance models are briefly presented, along with a description of their homoscedasticity and multiplicity control prerequisites. Following this, data collection and analysis procedures are detailed and their results reported and discussed.

Analytical models and statistical assumptions of variance

Analysis of variance refers to a collection of NHST models used in order to compare the differences among normal group means or medians in a sample. More precisely, ANOVA models test the null hypothesis that all group means or medians are equal by comparing the amount of systematic or explainable variance in the data to the amount of unsystematic or unexplainable variance. ANOVA models are thus omnibus tests: they tell us if the groups differ, but do not provide specific information about which groups differ and the magnitude of these multiple differences. In experimental terms, they allow us to determine if experimental manipulation has had some effect or not, but don't tell us specifically what the effect is.

Three fundamental types of ANOVA tests can be distinguished: parametric analysis of variance (ANOVA), parametric multivariate analysis of variance (MANOVA), and nonparametric analysis of variance. In the specific case of standard ANOVA tests, only one dependent variable is considered, and all models are variations on two common dimensions: how many independent variables they compare ('one-way' for one, 'factorial' for more than one, 'two-way' for two, ...) and what type of groups the test considers ('repeated measures' if the same participants are used in each variable group, 'independent' if the variable groups include different participants, and 'mixed' if both variable group types are used).

Multivariate analyses of variance (MANOVA) can be considered a generalization of ANOVA tests, as they can be used regardless of the number of independent variables involved, and are as such able to provide more information than univariate tests. "*If separate ANOVAs are conducted on each dependent variable, then any relationship between dependent variables is ignored. As such, we lose information about any correlations that might exist between the dependent variables. MANOVA, by including all dependent variables in the same analysis, takes account of the relationship between outcome variables. Related to this point, ANOVA can tell us only whether groups differ along a single dimension, whereas MANOVA (...) has greater power to detect an effect, because it can detect whether groups differ along a combination of variables*" (Field, 2012, p. 664).

Both standard ANOVA and MANOVA models are however parametric tests, in that they assume that the variables to be analyzed follow a given (normal) distribution. In contrast, nonparametric ANOVA models like the Kriskal-Wallis test obviate the need for distribution normality: by focusing on ranks and medians instead of absolute values and means, nonparametric tests represent an ingenious way to compare central tendencies between groups in cases where the data does not allow for parametric comparisons. Despite this versatility, nonparametric tests are however just as dependent as their parametric counterparts on their compliance with two fundamental statistical assumptions, without which no valid analysis of variance is possible: homoscedasticity and multiplicity.

Homoscedasticity

Like many other statistical tests, ANOVA models must comply to the rule that the groups whose means they compare are homoscedastic, that is, have similar variances (as opposed to heteroscedastic groups): "*as a rule of thumb, the largest and the smallest variance within groups should not differ by more than one order of magnitude. The reason for assuming equal residual variance (homoscedasticity) is that the variances are pooled to give an estimate of the variance that can be used for all groups simultaneously. (...) Slavish adherence to the formulae in the face of a violation of homoscedasticity assumption is one of the most common mistakes in the use of ANOVA which often result in a tragic abuse of really good data*" (Ståle, 1989, p. 270). Past research has indeed shown that violation of the homoscedasticity assumption for both ANOVA and MANOVA models has a nonnegligible impact on statistical significance; this has been observed in both unequal (Bishop, 1978; Box, 1954) and equal (Rogan, 1977) sample size contexts as well as under a wide variety of variance heterogeneity conditions (Bishop, 1976; Dajani, 2002; Krutchkoff, 1988; Lee, 2003; Lu, 2007). As for non-parametric tests, these are first and foremost statistical dominance tests, that is, procedures for identifying partial order between variables (Fishburn, 1964; Kmietowicz and Pearlman, 1979); as such, they are adequate for median difference testing only insofar as the groups being compared have similar shape and variance (Kasuya, 2001; Moser et al, 1989; Neuhäuser, 2002; Ogenstad, 1998).

Multiplicity

Multiplicity is for its part intrinsically linked to the scientific method itself. While science is often portrayed as a piecemeal investigative process, one which consists in asking nature "one question at a time" (Fischer, 1926, 511), actual scientific practice usually proceeds "familywise", by conducting large experiments designed to answer a collection of related claims. What is often forgotten however is that such batch testing and comparing increases the probability of detecting a nonexistent effect: for example, using a standard significance level of $\alpha = 5\%$ in a consecutive series of independent tests, one might expect to get a significant effect at the 20th test, regardless whether or not this effect is true (Bender, 2008). In addition, this inflation effect does not only concern the hypotheses of interest, but is instead tributary to all the potential inferences that can be drawn regarding the problems and data at hand. In other words, "familywise" statistical inferences are valid only insofar null hypotheses are scaled up accordingly. In the case of ANOVA, MANOVA, and nonparametric tests, this multiplicity problem emerges as soon as more than two variables are involved, one test being needed for each pairwise variable combination (Albers, 2019; Armitage, 1969; Cramer, 2016; Fletcher, 1989; John, 2012; Smith, 2002). As with NHST in general, the two most prudent and used statistical strategies to deal with false positives resulting from multiple comparisons are to control either the familywise error rate or the false discovery rate (Cramer, 2016). Both tests have proven their efficiency in a famous neuroimaging study, in which the authors showed that neglecting both tests enabled them to detect statistically significant signatures of brain activity in a dead Atlantic salmon during "mentalizing" tasks (Bennett, 2012). Despite this, researchers rarely use these tests: in a recent study of 819 articles published in six leading journals in psychology, it was shown that while almost half of all articles used a multiway ANOVA, only 1% of them used a correction procedure. Such a situation may stem from the fact that the multiplicity problem and its solutions are often not presented by mainstream statistical textbooks and statistical software packages, especially as regards to factorial, multivariate, and nonparametric tests (Cramer, 2016). Due to these omissions, not only are scholars not reminded to avoid considering the statistical significance of each mean comparison individually, but many might not even be aware of this necessity.

Methods

Article metadata for the corpus used in this research was obtained through the Web of Science (WoS) database, along with Scimago journal ranking statistics and the National Science Foundation disciplinary and sub-disciplinary classification of journals. Following this and with the use of various Python and R scripts, all WoS articles with digital object identifiers, published in 2015 in Elsevier journals from Life Sciences, Physical Sciences or Social Sciences & Humanities disciplines were kept. For each article of this data set, the full text was obtained using the Elsevier Full Text API, from which 321312 articles published amongst 1438 journals were extracted; WoS proportion and article frequencies by best quartile and discipline are shown in Figure 1. It is important to stress here that, while there are few Elsevier Journals in the lower quartiles of many disciplines, the total amount of articles within each discipline remains substantial and will allow for interesting analyses and comparisons.

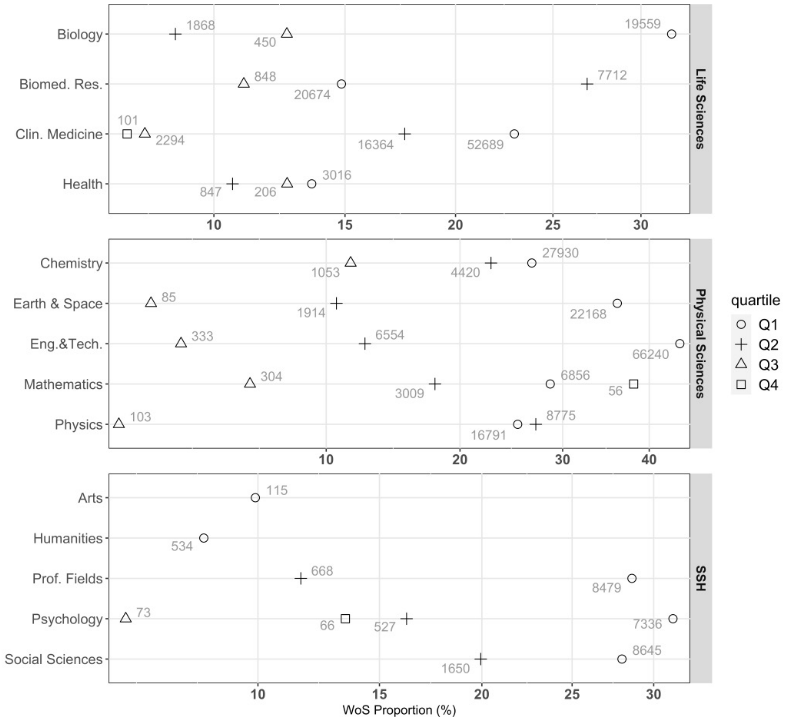


Figure 1. WoS proportion (%), along with article frequency of harvested Elsevier articles, by discipline and journal quartile.

In order to identify papers that use terms related to analysis of variance, homoscedasticity, and multiplicity, various regular expressions were designed and applied to the corpus. A first set of 3 regexes was aimed at identifying articles containing orthographic variants of expressions related to various analysis of variance tests. The first regular expression aims to capture all mentions of parametric analysis of variance in the corpus; negative lookaheads and lookbehinds were added to the regex in order to exclude all cases referring to MANOVA and non-parametric considerations. The second regex aims to capture references to multivariate or multiple analysis of variance (MANOVA). The third regular expression aims to catch expressions referring to the following non-parametric analysis of variance tests: Jonckheere's trend test, Friedman test, Kruskal-Wallis test, permutational ANOVA, non-parametric ANOVA as well as ANOVA on ranks. The last two regexes were designed to capture strings related to homoscedasticity and multiplicity considerations respectively. In the first case, references to either hetero- or homoscedasticity were considered relevant for the present study, hence the ulterior use of the expression 'scedasticity' to refer to such cases. This scedasticity regex was designed in order to capture variants of the following expressions: *homoscedasticity*, *heteroscedasticity*, *equal/similar/comparable variance*, *Levene's test*, *Bartlett's test*, *Hartley's fMax*, *Glejser test*, *White test*, *Cochran's C test*, *Box's M test*, *Breusch-Pagan test*, *Goldfeld-Quandt test*, *Brown-Forsythe test*, *Cook-Weisberg test*, *Harrison-McCabe test*, *Fligner-Killeen test*. As for multiplicity considerations, variants of the following strings were captured: *Bonferroni correction*, *family-wise error rate (FWER)*, *false discovery rate* (*FDR*), *Dunn's test, Dunnett's test*, *Šidák correction*, *Tukey's range test*, *Holm-Bonferroni step-down procedure*, and *Hochberg's step-up procedure*. In the case of ANOVA-related expressions referring to last names, regular expressions were designed in order to ignore citations; this was done by excluding all expressions followed by substrings characteristic of various citation formats, namely the abbreviation *et al.*, digits surrounded by brackets or parentheses as well as the conjunction *and* followed by a proper noun followed by digits surrounded by brackets or parentheses, as in *Smith (2008)* or *Johnson [3]*. Application of these regex to the corpus resulted in the extraction of 178818 expressions distributed over 55746 articles; frequency distribution of all captured strings for each regular expression are shown in the Appendix.

Results

The Venn diagram shown in Figure 2 classifies the articles containing strings captured by the above-mentioned regular expressions. Of the lot, 11 articles contain strings matched by all regular expressions, while 37043 articles match only one of the above regexes. Of particular interest for the present research are the set-theoretic equations shown below the figure: of the 13.6% of extracted papers containing parametric ANOVA mentions, 8.2% of them also contained expressions referring to scedasticity, 37.2% contained multiplicity-related expressions, and only 3.8% of them referred to both statistical assumptions. As regards to both MANOVA and nonparametric ANOVA test types, while the proportion of extracted papers is considerably smaller than in the case of ANOVA-related papers, proportions regarding statistical assumptions remain similar, with respectively 16.7% and 10.5% for scedasticity, 31.8% and 32.9% for multiplicity, as well as 6.4% and 5% for both assumptions. In sum, while around a third of all papers dealing with analysis of variance include considerations related to multiplicity correction, that ratio drops drastically in the case of scedasticity reporting, even more so in the case of papers alluding to both statistical assumptions. Finally, with the sole exception of multiplicity, proportions are higher in the MANOVA and nonparametric cases than in the parametric ANOVA one; this trend suggests that reliance on more sophisticated or marginal analysis of variance tests or considerations might be positively correlated to both the awareness of and commitment to the tests' underlying statistical assumptions.

Percentages for ANOVA-related articles (A), ANOVA-related articles that contain either scedasticity or multiplicity (SorM) references, and ANOVA-related articles that allude to both assumptions (SandM) were compiled for all extracted articles; results by disciplinary group, discipline, and journal quartile are shown in Figure 3. Overall, 14,3% of extracted Elsevier journal articles contain expressions referring to analysis of variance. Life Sciences papers have the highest proportion of such papers (27.7%), followed by Social Sciences & Humanities (17.3%) and Physical Sciences (3.6%), the latter containing more than half (51.8%) of all extracted papers. At the disciplinary level, Psychology (41.7%) and Biology (40.5%) have the highest proportion of ANOVA-related papers, while Physics (0.6%) and Mathematics (0.8%) have the lowest ratios. Given the disciplines involved and with the exception of Mathematics

Q4, it seems reasonable to conclude here that disciplines that rely more on NHST-based experimental procedures show a higher proportion of ANOVA-related papers.

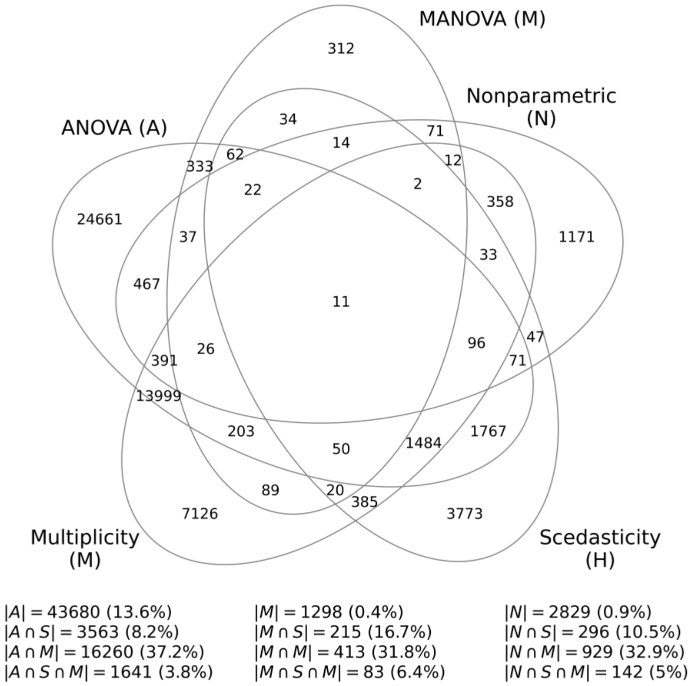


Figure 2. Article Frequency by Regex captured.

Regarding statistical assumptions, 39.7% of papers referring to analysis of variance include either scedasticity or multiplicity considerations; at the disciplinary group level, Life Sciences disciplines (42.4%) have on average a higher ratio than both SSH (34.1%) and Physical Sciences (28%), with Clinical Medicine (44.9%) and Biomedical Research (44.3%) leading the disciplinary landscape. Results are however markedly lower for papers alluding to both assumptions. Overall, only 3.6% of ANOVA-related papers refer to both scedasticity and multiplicity; proportions are uniformly distributed amongst disciplinary groups (Life Sciences: 3.4%, Physical Sciences: 4.4%, SSH: 3.6%) and disciplines, Mathematics Q3 (25%) and Earth & Space (8.7%) being the odd ones out in that respect. In sum, while the above Venn diagram showed that assumption testing was not thoroughly reported globally, analysis at both disciplinary group and discipline levels shows that this underreporting is not specific to a disciplinary group or a few disciplines, but rather pervades the whole scientific community, regardless of the subject matter and its formal complexity. It is also interesting to point out that higher-quartile journals tend to contain higher A, SorM, and SandM ratios than lower-quartile journals from the same discipline. While hinting at the appeal and impact of ANOVA-based research, this trend also suggests that more scrupulous peer-review in higher-quartile journals may help foster thorough statistical reporting in these publications.

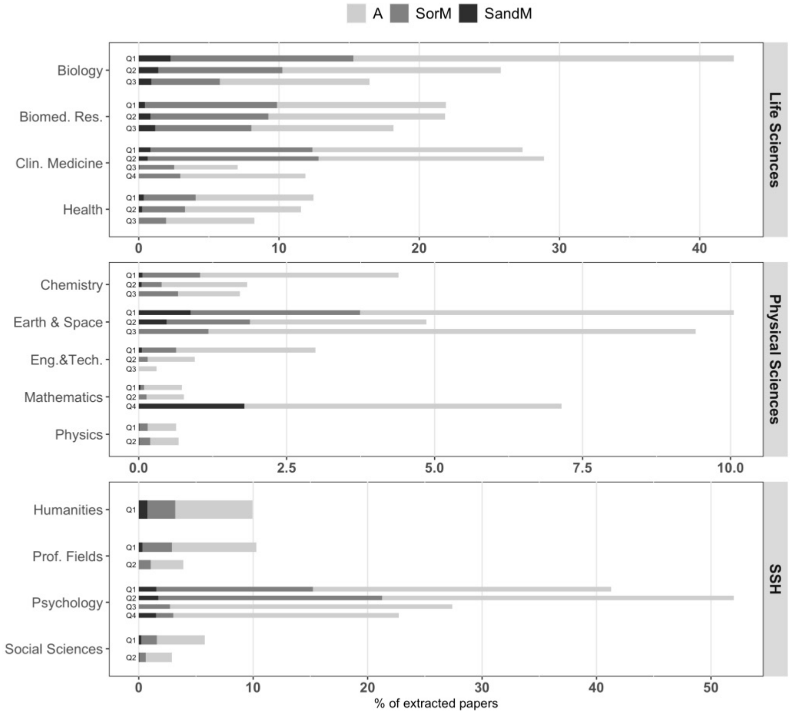


Figure 3. % of ANOVA-related papers (A) and ANOVA-related papers referring to either (SorX) or both (SandX) scedasticity and multiplicity, by discipline and journal quartile

At the finest granularity level, Figure 4 shows percentile proportions for the 5 subdisciplines of each disciplinary group that have the highest proportion of ANOVA-related papers. At first glance, the disciplinary composition of the group gives additional support to the hypothesized enmeshment of NHST-based experimental research and ANOVA testing. Also interesting is the fact that a few subdisciplines from Life Sciences and Psychology have a majority of papers referring to ANOVA models and tests, namely Physiology (67%), Experimental Psychology (59.2%), Behavioral Science & Complementary Psychology (52%) as well as Zoology (51.4%). As for assumption testing, Physiology (61.4%), Marine Biology & Hydrobiology (59.8%) as well as Pharmacology (53.9%) have a majority of ANOVA-related papers that include considerations referring to either assumption. However, ANOVA-related papers referring to both assumptions remain pretty low at the subdiscipline level, with perhaps the exception of Marine Biology & Hydrobiology (18.1%), Physiology (16.1%), Industrial Engineering (13.6%), and Oceanography & Limnology (12.8%). At the other end of the spectrum, journals from some subdisciplines and sub-disciplinary quartiles refer heavily to ANOVA, while scarcely mentioning anything about assumption tests; this is particularly true of Social Psychology, whose high proportion of ANOVA-related papers (37.9%) markedly contrasts with its relative silence as regards to assumption tests, with respectively 18.1% and 0.6% of ANOVA-related papers in that field containing references related to either and both statistical prerequisites. As for comparisons between subdiscipline quartiles, much variation can again be observed for the different ratios, and a pattern similar to the one observed at the discipline level seem to emerge from the data: papers published in higher-quartile journals tend to have higher ratios than papers published in lower-quartile journals. However, too many disciplinary quartiles are absent from the current dataset to make definitive statements in that regard. Overall and perhaps more than the previous analyses, these latter results show that, in the same manner as the use of ANOVA models itself, the reporting adequacy of these models' assumptions varies widely between subject matters, journal quartiles as well as between the very specific scientific subcultures that pertain to them.

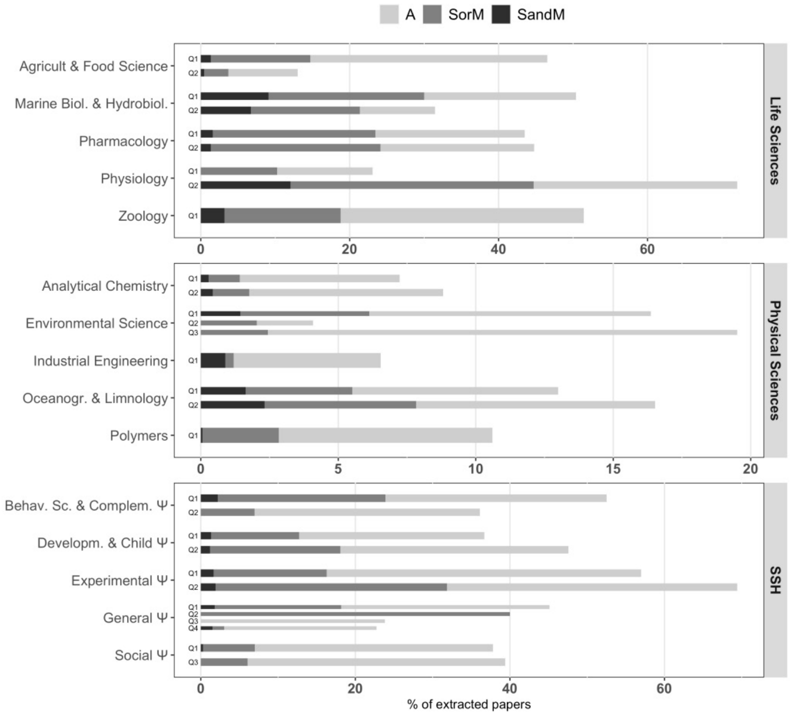


Figure 4. Percentage of ANOVA-related (A) papers and ANOVA-related papers referring to either (SorX) or both (SandX) scedasticity and multiplicity, by subdiscipline and journal quartile.

Conclusion

Overall, the present research shows that, beyond important variations at the different disciplinary levels, reporting transparency in the testing of ANOVA statistical assumptions is generally lacking, as a minority of ANOVA-related papers contain expressions referring to multiplicity, and even less to homo/heteroscedasticity or to both. Disciplinary groups, disciplines, and subdisciplines that have the highest proportion of ANOVA-related papers tend to be experimentation-prone and to refer more systematically to both statistical assumptions, with the notable exception however of social psychology. As regards to journal quartiles, a certain pattern can be observed at the different disciplinary levels, as articles from higher-quartile journals tend to report statistical assumptions more thoroughly than those from lower-quartile journals, regardless of discipline or subdiscipline; however, given that most Elsevier journals are to be found at the upper quartiles of each discipline or subdiscipline, more data is needed in the lower quartiles before drawing any conclusion in this regard. Despite these data gaps, the general picture presented here remains clear: efficient and thorough statistical reporting is still a long way ahead, being so far hinted at by only a few subdisciplines.

Of course, lack of reporting transparency does not any way imply lack of testing: it could be the case that in many non-reporting cases, statistical assumptions for the corresponding ANOVA models were actually tested. However, in the absence of any mention to that effect, nothing can be known about the completion of these procedures nor their outcome. In fact, non-testing is but one possible explanation of non-reporting; the possibility that tests were made but not reported due to negative or inconclusive results is also an equally if not more likely hypothesis, hence the importance of establishing thorough statistical reporting practices. On a related note, it might also be objected that non-testing is not necessarily indicative of malicious intent, but could also be attributed to lack of proper knowledge regarding statistics or their reporting. However, such a hypothesis is not only untestable, but also neglects the fact that, in matters of scientific integrity, practices that are the result of ignorance and negligence are indiscernible in their effect from those stemming from unhindered malevolence.

While these considerations and the results that trigger them are certainly grim, the dice are not yet cast in matters of statistical integrity. Indeed, the fact that so many authors from different and often unrelated fields explicitly comply to the testing of these statistical assumptions as well as the reporting of the latter's results suggests two things: first, that the scientific criticality of statistical assumption testing and reporting is acknowledged in practice by certain scientific groups; second, that collective awareness of and adherence to these issues is possible and can be fostered. On this matter, given both the inefficiency of guidelines in addressing statistical reporting issues (Clayson, 2019) and the high interquartile variability reported here, it could be argued that thorough peer reviewing would go a long way in rectifying the situation.

Acknowledgments

This work is supported by the Practices, Perceptions, and Patterns of Research Integrity (PRINT) project, funded by the Danish Agency for Higher Education and Science (Danish Ministry of Higher Education and Science) under grant No 6183-00001B. We would like to thank Pr. Jesper Wiborg Schneider, Serge Pascal Johannes M. Horbach, and Trine Theresa Holmberg for their guidance and support. *Cet article est dédié à la mémoire de Lucas.*

References

Albers, C. (2019). The problem with unadjusted multiple and sequential statistical testing. *Nature communications*, 10, 1, 1-4.

Armitage, P., McPherson, C. & Rowe, B. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, 132(2), 235–244.

Bender, R., Bunce, C., Clarke, M., Gates, S., Lange, S., Pace, N.L. & Thorlund, K. (2008). Attention should be given to multiplicity issues in systematic reviews. *Journal of clinical epidemiology,* 61(9), 857-865.

Bennett, C.M., Baird, A.A., Miller, M.B. & Wolford, G.L. (2012). Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for proper multiple comparisons correction. *Journal of Serendipitous and Unexpected Results*, 1(1), 1–5.

Bishop, T.A. & Dudewicz, E.J. (1978). Exact analysis of variance with unequal variances: test procedures and tables. *Technometrics*, 20(4), 419–430.

Bishop, T.A. (1976). *Heteroscedastic ANOVA, MANOVA, and multiple-comparisons* (Doctoral dissertation, The Ohio State University).

Box, G.E. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. *The Annals of Mathematical Statistics*, 25(3), 484-498.

Chalmers, I. (1990). Underreporting Research Is Scientific Misconduct. *Journal of the American Medical Association*, 263(10), 1405-1408.

Clayson, P.E., Carbine, K.A., Baldwin, S.A. & Larson, M.J. (2019). Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. *Psychophysiology*, 56(1), e13437.

Cramer, A.O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R.P., Waldorp L.J. & Wagenmakers, E.J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, 23(2), 640-647.

Dajani, A.N. (2002). *Contributions to statistical inference for some fixed and random models* (Doctoral dissertation, University of Maryland Baltimore County).

Fiedler, K. & Schwarz N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, 7(1), 45-52.

Fishburn, P.C. (1964). *Decision and Value Theory*. Wiley, New York.

Fletcher, H.J., Daw, H. & Young, J. (1989). Controlling multiple F test errors with an overall F test. *The Journal of Applied Behavioral Science*, 25(1), 101-108.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5), 587-606.

Hardwicke, T.E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S.N. & Ioannidis, J.P. (2020). Calibrating the scientific ecosystem through meta-research. *Annual Review of Statistics and Its Application*, 7, 11-37.

John, L.K., Loewenstein, G. & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524-532.

Kasuya, E. (2001). Mann-Whitney U test when variances are unequal. *Animal Behaviour*, 61, 1247-1249.

Kmietowicz, Z.W. & Pearman, A.D. (1979). Decision Theory and Weak Statistical Dominance. *Journal of Operational Research Society*, 30(11), 1019–1022.

Krutchkoff, R.G. (1988). One-way fixed effects analysis of variance when the error varianes may be unequal. *Journal of Statistical Computation and Simulation*, 30(4), 259-271.

Lee, S. & Ahn, C.H. (2003). Modified ANOVA for unequal variances. *Communications in Statistics-Simulation and Computation*, 32(4), 987-1004.

Leggett, N.C., Thomas, N.A., Loetscher, T. & Nicholls, M.E. (2013). The life of p: “Just significant” results are on the rise. *Quarterly Journal of Experimental Psychology*, 66(12), 2303-2309.

Lu, F. (2007). ANOVA and MANOVA under heteroscedasticity. (Doctoral dissertation, University of Louisiana at Lafayette, Lafayette, LA).

Martinson, B.C., Anderson, M.S. & De Vries, R. (2005). Scientists behaving badly. *Nature*, 435(7043), 737-738.

Moser, B.K., Stevens, G.R. & Watts, C.L. (1989). The two-sample t test versus Satterthwaite’s approximate F test. *Communications in Statistics: Theory and Methods*, 18, 3963–3975.

Neuhauser, M. (2002). Two-sample tests when variances are unequal. *Animal Behaviour*, 63(4), 823-825.

Ogenstad, S. (1998). The use of generalized tests in medical research. *Journal of Biopharmaceutical Statistics*, 8, 497-508.

Rogan, J.C. & Keselman, H.J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. *American Educational Research Journal*, 14(4), 493-498.

Sacco, D.F., Bruton, S.V. & Brown, M. (2018). In defense of the questionable: defining the basis of research scientists’ engagement in questionable research practices. *Journal of Empirical Research on Human Research Ethics*, 13(1), 45-52.

Schneider, J.W. (2015). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics*, 102(1), 411-432.

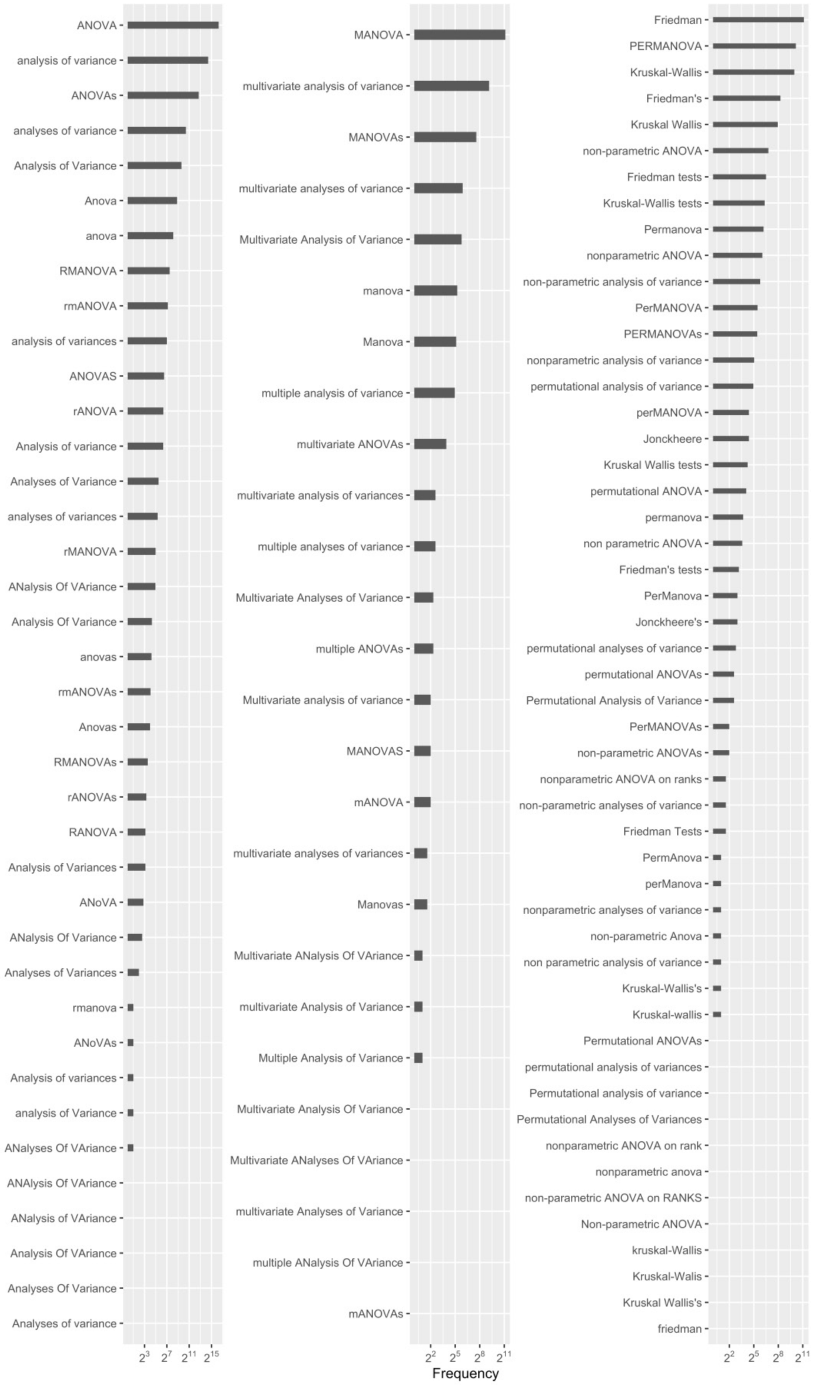
Smith, R.A., Levine, T.R., Lachlan, K.A. & Fediuk, T.A. (2002). The high cost of complexity in experimental design and data analysis: Type I and type II error rates in multiway ANOVA. *Human Communication Research*, 28(4), 515-530.

Steneck, N.H. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and Engineering Ethics*, 12(1), 53-74.

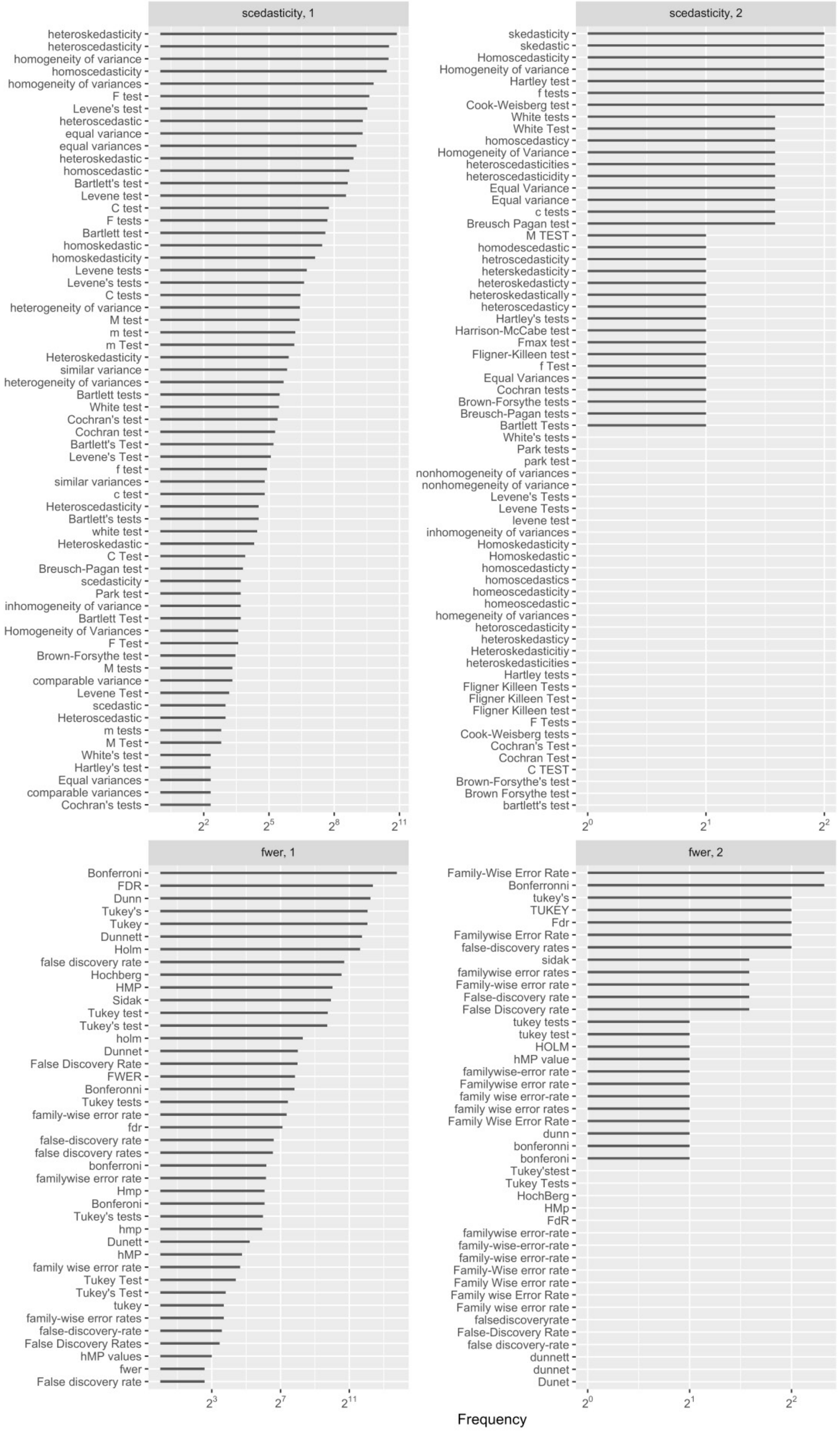
Ståle, L. & Wold, S. (1989). Analysis of variance (anova). *Chemometrics and Intelligent Laboratory Systems*, 6(4), 259-272.

Appendix: Regular expression string capture stats

ANOVA tests



Assumptions tests



Evaluating the impacts of international cooperation under institutional agreements involving funding agencies and research organizations

Sergio Salles-Filho[1], Paulo Feitosa[2], Adriana Bin[3] and Fernando Colugnati[4]

[1] *sallesfi@unicamp.br*
Department of Science and Technology Policy, University of Campinas, Campinas, SP (Brazil)

[2] *pfeitosa@usp.br*
Department of Public Relations, Advertising and Tourism, University of Sao Paulo, Sao Paulo, SP (Brazil)

[3] *adriana.bin@fca.unicamp.br*
School of Applied Sciences, University of Campinas, Campinas, SP (Brazil)

[4] *fernando.colugnati@medicina.ufjf.br*
Medical School, Federal University of Juiz de Fora, Juiz de Fora, MG (Brazil)

Abstract

Funding agencies (FAs) have increasingly engaged in international cooperation agreements (ICAs) to encourage world-class research and achieve more promising outcomes in the context of increasing competition for research resources. The main research question we want to address is the influence of institutionalized ICA (meaning between funding agencies and universities, and other research organizations) on scientific and technological outputs. Our research design is based on a comparison between granted researchers under ICAs (treatment group) and granted researchers not carried under ICAs (control group). Using a quasi-experimental evaluation design, we show that ICAs have a positive and significant impact on the quality of scientific production and the number of national and international collaborations. Different results were found in terms of the number of scholarly and technological outputs.

Introduction

Funding agencies (FAs) have played an irreplaceable role in promoting the generation and diffusion of scientific knowledge when examining recent history[1]. The performance of these organizations has been observed in terms of conventional research outputs, such as citation impact of publications (Yan et al. 2018), but also in its ability to modify the rate and direction of inventive activity (Corredoira et al., 2018). Positive effects are widely reported in the literature, although variations in efficiency between different funding systems are observed due to factors such as research evaluation system, competition for granted projects, and institutional autonomy (Aghion et al., 2010; Sandström and Van den Besselaar, 2018).

Given the need to enhance the results from funded research, FAs have increasingly engaged in international cooperation agreements (ICAs) (Reale et al., 2012). These agreements aim to encourage world-class research and achieve more promising outcomes in increasing competition for research resources. Moreover, in a world where scientific discovery increasingly relies on wide-spread connections, the agreements reduce institutional distances that represent obstacles to interaction among those involved in academic research (Sergi et al., 2014).

Arguments in favour of the engagement of FAs in institutional agreements with foreign research organizations could be relying on two main rationales. First, the quality and quantity of funded research may be enhanced when they are undertaken in collaboration, which is the mechanism that allows the exchange of ideas required for generating new knowledge, solving complex problems, and enabling innovative practices (Sergi et al., 2014). This rationale is adequately discussed in the literature since the study on research collaboration, including international

[1] These organizations are often quasi-public and "financed by the State in order to define and execute a large part of the science policy" (Braun, 1998).

collaboration, has received increasing attention from the scientific community, research institutions, and policymakers (Bammer, 2008; Narin et al., 1991; Wagner and Leydesdorff, 2005). More recently, the "International Research Collaboration" has been perceived as an emerging area of innovation studies (Chen et al., 2019). This research agenda covers topics such as the effects of international collaboration on research outputs (Bozeman et al., 2013; Lee and Bozeman, 2005), the role of international knowledge networks (Adams, 2013; Guan et al., 2017; Zhao and Guan, 2011) and collaboration patterns (Leydesdorff and Wagner, 2008; Todeva and Knoke, 2005; Wagner and Leydesdorff, 2005).

The second rationale is based on a relevant causal argument that a better outcome would not have occurred in the absence of the cooperation agreements. A pertinent question in the counterfactual language is: "Would the researchers be less collaborative or productive if they had collaborated out of institutional agreements?". The counterfactual evaluation of the magnitude of this specified causal effect is scarce. Yet, it is usually assumed among policymakers and scientists that FAs should be encouraged to engage in numerous agreements. The use of counterfactuals in policies has been the subject of disagreement among scholars; however, many issues for which causal inference would be feasible and useful.

This article contributes to the literature by evaluating the impacts of research granted under ICAs on scientific and technological outputs and the intensity of collaboration. Particularly, we employ a panel of researchers granted by The São Paulo Research Foundation (FAPESP) to compare those carrying collaboration under and out of ICA. Accordingly, we designed a quasi-experimental evaluation with samples of researchers granted by the Foundation between 1990 and 2018.

Literature review

The literature allows us to identify two distinct and complementary research streams. The first pay attention to the relationship between collaboration – often measured through the proxy of co-authorship (especially international co-authorship) – and research outputs measured by the number of papers and citations. The positive correlation between these variables has been confirmed by many studies (Glänzel and Schubert, 2001; He et al., 2009; Katz and Hicks, 1997; Moya-Anegon et al., 2018; Narin et al., 1991; Zhou and Bornmann, 2014). Furthermore, the comprehensive study by Wuchty et al. (2007) revealed that research produced by teams is frequently more cited than those that individuals do, and this prevalence has been increasing over time.

Nevertheless, the effect of international collaboration on increasing the citation impact of publications tends to vary significantly when analysed at the national level and depends mainly on the partner country's scientific development phase (Moya-Anegon et al., 2018). For instance, collaboration with the US increases the research impact, mostly when US researchers serve as corresponding authors (Bote et al., 2013; Lancho-Barrantes et al., 2013; Sud and Thelwall, 2016). Despite the positive effects of research produced through international collaboration in citations, Wagner et al. (2019) also reveal that international collaboration seems to have less novel and more conventional knowledge combinations.

The second group of studies emphasizes the relationship between funding and research collaboration and reveals that these dimensions interact with each other since funding enables international collaboration (Cimini et al., 2016; Clark and Llorens, 2012; Liu et al., 2018; Ubfal and Maffioli, 2011), and international collaboration facilitates access to funding (Zhou and Tian, 2014). Zhou et al. (2020) argue that this interaction produces a more complex effect on citation impact.

In terms of evidence, a study using a panel of top US universities' departments found that scientists who have earned prestigious awards tend to collaborate with larger research teams (Adams et al. 2005). The panel with scientists from the European Union reveals that the impact of collaboration on productivity is positive and significant (Defazio et al., 2009). The study also proposes that future research should include a control group of researchers applying for the same fund but not selected.

Further evidence shows that research grants in the United States have a substantial and significant impact on the number of collaborators affiliated with university research centres, although the net impacts of collaboration are less clear (Bozeman and Corley, 2004; Lee Bozeman, 2005). Other studies on the Argentine experience reveal that grants have a positive effect on the quantity and quality of the publications (Chudnovsky et al., 2008), and also impacts on collaboration, which is measured in terms of the number of co-authors for publications in peer-reviewed journals (Ubfal and Maffioli, 2011). These arguments lead us to conjecture:

Hypothesis 1. *The impact of the scientific outputs of granted researchers associated with ICAs is greater than granted researchers with international activity but not associated with ICAs.*

Hypothesis 2. *The impact of the technological outputs of granted researchers associated with ICAs is greater than granted researchers with international activity but not associated with ICAs.*

Hypothesis 3. *Granted researchers associated with ICAs have more scientific production in collaboration than granted researchers with international activity but not associated with ICAs.*

Methods

We use a unique dataset of 11.787 granted researchers, representing all FAPESP research grants between 1990 and 2018. These researchers were divided into two groups: granted researchers associated with ICAs (treatment) and granted researchers not related to ICAs but with some form of international collaboration (control). Therefore, from a population of 11,787 granted researchers, 572 are associated with ICAs, and 2,055 are not, but had carried various sorts of point-to-point collaboration. Table 1 – Describes the secondary data sources that supported this study.

Table 1 – Description of the databases.

Data	*Description*	*Source*
Research grant data	Information on research grants at FAPESP	FAPESP
CV data	Information from academic curricula registered on the Lattes CV platform	CNPq
Scientific production	Information on scientific production from the Lattes CV platform	CNPq
Scientific production	Information on scientific production from the Dimensions platform	Dimensions
Citations	Information on citation of scientific production from the Dimensions platform	Dimension
Technological production	Information on patent applications and utility model	INPI
Scientific production	Information on scientific production from the Scopus platform	Scopus
Citations	Information on citation of scientific production from the Scopus platform	Scopus

Note: Lattes Platform is a hugely comprehensive and open CV database in Brazil; CNPq: Brazilian National Science Foundation; INPI, National Institute for Intellectual Property.

Although some authors refer to this as a quasi-experimental approach (Ferraro and Pattanayak, 2006) we rather take it as a comparison group using estimation of Average Treatment on the

Treated (ATT) effects, likely less biased than traditional ATE in this type of selection process. This approach is particularly necessary when one intends to measure the funding instrument's difference (Ferraro, 2009; Ferraro and Pattanayak, 2006; Frondel and Schimidt, 2005). Quasi-experiments require the consideration of four main aspects (Ferraro; Pattanayak, 2006), namely: 1) consider that the effects are the same for both groups; 2) conjecture what are the potential effects resulting from the intervention (of the program); 3) design a simple control group (those who did not receive the intervention) and 4) collect data on the effects and key inputs before (baseline) and after the intervention.

The evaluation adopted the Coarsened Exact Matching (CEM) method, which proved to be adequate compared to Exact and Nearest Neighbour. The observable baseline variables used in the pairing were: Year of first FAPESP research grant not related to ICAs; Year of the first FAPESP research grant associated to ICAs; time-period of the project under analysis; Responsible institution; Areas of knowledge; and Total FAPESP research grant. Variables as for number of publications and citations before the period of comparison were adjusted. After matching, the balance reduced the heterogeneity between the groups, and the Standardized Average Differences practically zeroed after matching, maintaining a sufficient sample size for inferences.

Results

The estimates of impacts on scientific, technological, and collaboration outputs are presented in tables 2, 3, and 4, respectively.

Table 2 – Impact on scientific outputs

	Effect	*p-value*	*CI 95 %*
Number of articles (Lattes)	0.927	0.432	[0.766 – 1.121]
Number of articles (Dimensions)	1.008	0.944	[0.813 – 1.249]
Number of citation (Dimensions)	1.041	0.889	[0.595 – 1.821]
Number of citation (Scopus)	1.964	<0.001	[1.468 – 2.628]
H-index (Scopus)	1.181	0.008	[1.045 – 1.335]

Note: Number of articles and citations in the 5 years before and after the first grant.

Table 3 – Impact on technological outputs

	Effect	*p-value*	*CI 95 %*
Number of Patents (INPI)	0.582	0.293	[0.213 – 1.594]

Table 4 – Impact on national and international collaboration

	Effect	*p value*	*CI 95 %*
Single authorship (Scopus)	1.096	0.651	[0.736 – 1.633]
International collaboration (Scopus)	1.567	<0.001	[1.245 – 1.973]
National collaboration (Scopus)	1.301	0.019	[1.045 – 1.619]

The main findings can be summarized as follows:

a. It **was not possible** to state that ICAs produce a quantitative impact on scientific production when comparing the number of articles in journals.
b. It **was possible** to conclude that ICAs make a qualitative difference in scientific production based on citations and H index: 96% more citations and H index 18% higher in the treatment group.
c. It **was not possible** to state that ICAs produce higher technological production when comparing the number of patents.

d. It **was possible** to confirm that collaboration under ICAs produces higher international collaboration (by 56%). This cooperation effect also extends to national cooperation, which is 30% higher in the treatment group. In other words, researchers participating in ICAs are more cooperative than those who collaborate internationally outside institutional agreements.

References

Adams, J. 2013. The fourth age of research, *Nature*, vol. 497, no. 7451, 557–60

Aghion, P., Dewatripont, M., Hoxby, C., Mas-Colell, A., and Sapir, A. 2010. The governance and performance of universities: evidence from Europe and the US, *Economic Policy*, vol. 25, no. 61, 7–59

Bammer, G. 2008. Enhancing research collaborations: Three key management challenges, *Research Policy*, vol. 37, no. 5, 875–87

Bote, V. P. G., Olmeda-Gómez, C., and Moya-Anegón, F. de. 2013. Quantifying the Benefits of International Scientific Collaboration, *Journal of the American Society for Information Science and Technology*, vol. 64(2), no. July, 392–404

Bozeman, B. and Corley, E. 2004. Scientists' collaboration strategies: Implications for scientific and technical human capital, *Research Policy*, vol. 33, no. 4, 599–616

Bozeman, B., Fay, D., and Slade, C. P. 2013. *Research collaboration in universities and academic entrepreneurship: The-state-of-the-art*

Braun, D. 1998. The role of funding agencies in the cognitive development of science, *Research Policy*, vol. 27, no. 8, 807–21

Chen, K., Zhang, Y., and Fu, X. 2019. International research collaboration: An emerging domain of innovation studies? *Research Policy*, vol. 48, no. 1, 149–68

Chudnovsky, D., López, A., Rossi, M. A., and Ubfal, D. 2008. Money for science? The impact of research grants on academic output, *Fiscal Studies*, vol. 29, no. 1, 75–87

Cimini, G., Zaccaria, A., and Gabrielli, A. 2016. Investigating the interplay between fundamentals of national research systems: Performance, investments, and international collaborations, *Journal of Informetrics*, vol. 10, no. 1, 200–211

Clark, B. Y., and Llorens, J. J. 2012. Investments in Scientific Research: Examining the Funding Threshold Effects on Scientific Collaboration and Variation by Academic Discipline, *Policy Studies Journal*, vol. 40, no. 4, 698–729

Corredoira, R. A., Goldfarb, B. D., and Shi, Y. 2018. Federal funding and the rate and direction of inventive activity, *Research Policy*, vol. 47, no. 9, 1777–1800

Defazio, D., Lockett, A., and Wright, M. 2009. Funding incentives, collaborative dynamics, and scientific productivity: Evidence from the EU framework program, *Research Policy*, vol. 38, no. 2, 293–305

Domingues, A. A. and Costa, M. C. da. 2016. A colaboração internacional da FAPESP: quais contextos para suas transformações?, *Revista Brasileira de História da Ciência*, vol. 9, no. 1, 19–35

Glänzel, W. and Schubert, A. 2001. Double effort = Double impact? A critical view at international co-authorship in chemistry, *Scientometrics*, vol. 50, no. 2, 199–214

Guan, J., Yan, Y., and Zhang, J. J. 2017. The impact of collaboration and knowledge networks on citations, *Journal of Informetrics*, vol. 11, no. 2, 407–22

He, Z. L., Geng, X. S., and Campbell-Hunt, C. 2009. Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a New Zealand university, *Research Policy*, vol. 38, no. 2, 306–17

Katz, J. S., and Hicks, D. 1997. How much is a collaboration worth? A calibrated bibliometric model, *Scientometrics*, vol. 40, no. 3, 541–54

Lancho-Barrantes, B. S., Guerrero-Bote, V. P., and de Moya-Anegón, F. 2013. Citation increments between collaborating countries, *Scientometrics*, vol. 94, no. 3, 817–31

Lee, S. and Bozeman, B. 2005. The impact of research collaboration on scientific productivity, *Social Studies of Science*, vol. 35, no. 5, 673–702

Leydesdorff, L. and Wagner, C. S. 2008. International collaboration in science and the formation of a core group, *Journal of Informetrics*, vol. 2, no. 4, 317–25

Liu, A. M. M., Liang, O. X., Tuuli, M., and Chan, I. 2018. Role of government funding in fostering collaboration between knowledge-based organizations: Evidence from the solar PV industry in China, *Energy Exploration and Exploitation*, vol. 36, no. 3, 509–34

Moya-Anegon, F. de, Guerrero-Bote, V. P., Lopez-Illescas, C., and Moed, H. F. 2018. Statistical relationships between corresponding authorship, international co-authorship and citation impact of national research systems, *Journal of Informetrics*, vol. 12, 1251–62

Narin, F., Stevens, K., and Whitlow, E. S. 1991. Scientific cooperation in Europe and the citation of multi nationally authored papers, *Scientometrics volume*, vol. 21, no. 3, 313–23

Reale, E., Inzelt, A., Lepori, B., and Van Den Besselaar, P. 2012. The social construction of indicators for evaluation: Internationalization of Funding Agencies, *Research Evaluation*, vol. 21, no. 4, 245–56

Sandström, U. and Van den Besselaar, P. 2018. Funding, evaluation, and the performance of national research systems, *Journal of Informetrics*, vol. 12, no. 1, 365–84

Sergi, B., Parker, R., and Zuckerman, B. 2014. Support for international collaboration in research: The role of the overseas offices of basic science funders, *Review of Policy Research*, vol. 31, no. 5, 430–53

Sud, P. and Thelwall, M. 2016. Not all international collaboration is beneficial: The Mendeley readership and citation impact of biochemical research collaboration, *Journal of the American Society for Information Science and Technology*, vol. 67(8), 1849–57

Todeva, E. and Knoke, D. 2005. Strategic alliances and models of collaboration, *Management Decision*, vol. 43, no. 1, 123–48

Ubfal, D. and Maffioli, A. 2011. The impact of funding on research collaboration: Evidence from a developing country, *Research Policy*, vol. 40, no. 9, 1269–79

Wagner, C. S., and Leydesdorff, L. 2005. Network structure, self-organization, and the growth of international collaboration in science, *Research Policy*, vol. 34, no. 10, 1608–18

Wagner, C. S., Whetsell, T. A., and Mukherjee, S. 2019. International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination, *Research Policy*, vol. 48, no. 5, 1260–70

Wuchty, S., Jones, B. F., and Uzzi, B. 2007. The Increasing Dominance of Teams in Production of Knowledge, *Science*, vol. 316, no. 5827, 1036–39

Zhao, Q., and Guan, J. 2011. International collaboration of three 'giants' with the G7 countries in emerging nanobiopharmaceuticals, *Scientometrics*, vol. 87, no. 1, 159–70

Zhou, P. and Bornmann, L. 2014. An overview of academic publishing and collaboration between China and Germany, *Scientometrics*, vol. 102, no. 2, 1781–93

Zhou, P., Cai, X., and Lyu, X. 2020. An in-depth analysis of government funding and international collaboration in scientific research, *Scientometrics*, vol. 125, no. 2, 1331–47

Zhou, P. and Tian, H. 2014. Funded collaboration research in mathematics in China, *Scientometrics*, vol. 99, no. 3, 695–715

The small world of editorships: A network on innovation studies

Ana Teresa Santos[1] and Sandro Mendonça[2]

[1] *atmss@iscte-iul.pt*
Instituto Universitário de Lisboa (ISCTE-IUL) – Business Research Unit (BRU-IUL), Lisbon, Portugal.

[2] *sfm@iscte-iul.pt*
Instituto Universitário de Lisboa (ISCTE-IUL) – Business Research Unit (BRU-IUL), Lisbon, Portugal.
UECE/REM – ISEG/ University of Lisbon, Lisbon, Portugal.
SPRU, University of Sussex, UK.

Abstract

Editors exert a significant influence on journal's mission and governing the strategic direction of outlets. They are the channels gatekeepers not only by ensuring the quality but also by guaranteeing the integrity of novels produced. For being such an important piece of scientific puzzle, they are a research object of utmost interest which is rather fragmented. This paper aims to better understand the relationships between editors seated on boards of 20 innovation top-tiers. The sample considered comprised 2,440 editors occupying 3,005 editorial positions and assuming 122 different duties. No single journal is free from this interlocking editorship phenomenon and 18.6% of the scholars serve on multiple boards. We deploy social network analysis to further inquire and model the editorial relationships in which innovation journals are embedded. Our results offer new insights on how the field is organised: 627 lines linking the journals were found with a 41.6% interlocking density. Research Policy has the highest number of direct links to other boards (degree) and the shortest distance from all network journals (closeness) while Industrial and Corporate Change is the one bridging the largest number of other pairs of journals (betweenness), followed by Small Business Economics and Research Policy.

Introduction

Elite boards memberships are crucial agents in scientific governance. For being seen as critical the role they play and thus, an appointment to a journal board is considered an important career stepping-stone as it provides opportunities for intellectual growth and networking (Topaz & Sen, 2016). The critical mentality and decisions of scientific editors have so far safeguarded and will also warrant the social and intellectual integrity of science for the upcoming years. For being consulted about research agendas and strategic directions for the outlets, the elite board membership become of paramount interest (Bedeian, Van Fleet, & Hyman, 2009; Feldman, 2008). Their positions and editorial affiliations provide a chance to study the underlying direction of journals (Morton & Sonnad, 2007; Wilkes & Kravitz, 1995).

As prominent scholars with a robust track of publications (Teixeira & Oliveira, 2018) and highly appreciated by peers (Andrikopoulos & Economou, 2015), editors are commonly seat in more than one board. Such phenomenon was previously identified by Baccini & Barabesi (2010) and may be responsible for the establishment of subgroups of scholars linked to some core journals, who may exert influence on the vision and main paradigms of such journals.

Taking advantage of commonly described academic boards on journal's website, we aim to ponder about journal governance as already done for other scientific fields (Bakker & Rigter, 1985; Brinn & Jones, 2008; Burgess & Shaw, 2010).

In this work, we draw a social network analysis to discuss about the social structures and independence of journal's EB. Through the assessment of Boards composition of the twenty most important innovation journals identified by Fagerberg et al. (2012), we examined how the memberships of EB interact and identified the most influential ones in the field. We believe this study may introduce some pertinent knowledge for those interested social structures in the innovation studies context.

The editorship network

In modern science governance literature, editors validate their role legitimacy through high academic standing and further signal their intellectual and social capital resources through board member affiliations. In this sense, the editorial process becomes an important professional network. To investigate the relationships between editors and journals, we have employed principles from network science to study complex systems composed of relationships between entities (Vespignani, 2018).

A social network was defined by Wasserman and Faust (1994), p.20, "as a finite set or sets of actors and the relationship or relationships between them". With social network analysis (SNA), we can find groups of elite board memberships surrounded scientific outlets or as bridges connecting them. Social network analysis characterizes networked structures in terms of nodes and ties (edges). Networks can be conceptualized organizationally, as networks of journals connected by editors. In this work, we model the relation between editors from innovation-oriented journal Boards based on data collected on outlets' webpages.

Considering editors seating on more than one journal Board is a proxy of intellectual similarity between editorial policies, we may perceive journals have closer policies according to the number of scholars they have in common on their Boards. The interlocking phenomenon puts outlets closer to each other and facilitates the communication. In other words, the closeness of the editorial policies of two scientific journals can be assessed by the number of common editors sitting on their Boards. We will not focus on the editorial policies adopted by the Boards of innovation-oriented journals. Instead, we will infer about the similarity of editorial policies through the detection of recurrent scholars as common editors between Boards.

Exploring editorial teams in "Innovation Studies" periodicals

Innovation studies is an evolving interdisciplinary field focused on producing systematic and reliable knowledge about how best to influence innovation and to exploit its effects to the full (Fagerberg, Martin, & Andersen, 2013). Born from plural contexts like Economics (Nelson, 1959), Management (Burns & Stalker, 1961) and Sociology (Rogers, 1983), it became a global research community world-wide (Martin, 2012). Fagerberg et al. (2012) analysed the development of innovation studies and through an empirical approach based on analysing the chapters contained in authoritative handbooks on innovation studies, identified which publications had most impact (Fagerberg and Verspagen, 2009). This proved to be consistent with later studies from Cancino, Merigó, & Coronado, (2017), Kotsemir (2013) and Rakas and Hain (2019). With the purpose of understanding the editorial community, we studied the emergence of the innovation studies field from an editorial point of view (Fagerberg, Mowery, & Nelson, 2004). In this study, we restricted our analyses to the twenty most influential journals identified (Fagerberg et al., 2012).

Our research explores structural properties of the network generated by the editorial population of leading innovation studies journals (de Andrade & Rêgo, 2018). The general aim of this study is to investigate the relationships between editors and journals. Taking advantage of centrality measures such as degree, betweenness and closeness, the most central outlets and their roles in the network were also identified.

The Boards of "Innovation Studies"

Data collection and methodology

From outlets' editorial pages, we collected scholars' names, institutional affiliations, gender and their roles inside the Board. A summary about the editorial memberships found is provided in Table 1, including the number of editorial positions available, the number of different scholars seating on the Board and shared editors with other outlets. In total, 2,440 different persons were found for the 3,005 editorial positions available. Repeated names allowed the identified the ones in charge of multiple duties inside a journal or among different journals.

Table 1. Editorial Boards descriptive characteristics.

Short name	Journal	No. of editorial memberships	Total distinct scholars	Shared editors	No. of duties
AMJ	Academy of Management Journal	328	328	145	5
AMR	Academy of Management Review	312	310	149	5
ASQ	Administrative Science Quarterly	115	111	62	5
CJE	Cambridge Journal of Economics	53	52	1	4
HR	Human Relations	99	99	26	1
ICC	Industrial and Corporate Change	98	98	38	8
IJTM	International Journal of Technology Management	21	21	2	3
JIBS	Journal of International Business Studies	275	274	66	8
JMS	Journal of Management Studies	280	280	99	2
MS	Management Science	399	365	21	35
OSc	Organization Science	237	237	117	4
OSt	Organization Studies	235	234	82	6
RDM	R&D Management	19	19	3	5
RS	Regional Studies	37	37	3	11
RP	Research Policy	102	102	38	3
SBE	Small Business Economics	152	150	14	4
SMJ	Strategic Management Journal	50	50	28	6
TASM	Technology Analysis & Strategic Management	37	37	12	3
TFSC	Technological Forecasting and Social Change	98	96	12	5
Tec	Technovation	58	58	12	12

Source: Scimago, as of March 2019 and journals' homepage.

Among those top-tier outlets, it was noticed the editorial memberships available differ greatly between journals. Some of them revealed to have small numbers of editorial positions such as RDM and IJTM while other Boards reported more then 300 of duties entrusted to scholars. For 12 outlets, these duties were assigned to different personalities while 8 top-tiers made some editors responsible for multiple roles. All outlets also share at least one editor with other Board, actually, scholars shared range from 2% to 56%. For this interest, duties and the proportion of shared editors is addressed apart.

Editors' duties

Among different journals, diverse internal organisations within EB were found. There are journals with only one title for all editors, such as observed for HR where everyone is “editor” while MS exhibited 35 different titles within the Board. According to the titles assigned, different internal organisations were supposed: journals like HR, where the same title label is given to all editors without further hierarchies contrasted with other Boards which presented more defined internal structures with five or more different categories.

It is noteworthy that outlets have different numbers of scholars on Board. With the exception of MS, all journals have one title ascribed to a majority of the editors. Editorial duties' labels were kept exactly as recorded from official journals' webpages with exception of plural descriptions which were converted to singular. The lack of uniformity in the similar positions' titles across journals makes it harder to compare editors' responsibilities. As expected from Table 1, MS counts with a large number of memberships while RDM has a very small editorial team. ICC and IJTM have also a geographical organisation for editors on Board which is not found in the other outlets.

Although AMJ and AMR exhibit a similar editorial structure, with a team encompassing mainly scholars as part of Editorial Board, in general, outlets from the same publisher did not present a similar editorial organisation. Considering the ones published by John Wiley & Sons (JMS, RDM and SMJ), it is clear the heterogeneous labels used for the group of scholars on their Boards. The main editorial assignment in JMS, named as “Editorial Board”, includes 279 scholars while RDM has five different categories and the largest one, “Editorial Advisory Board”, has only eleven memberships. SMJ organises editors among six different categories involving 33 as “Associate Editor”.

The same editorial title is also found between Boards with different frequencies, suggesting dissimilar commitments. In HR, those 99 scholars on its Board are assigned as “Editor” and no other group of editors are disclosed. However, in RDM, the one assigned as “Editor” seems to be the main gatekeeper in this top-tier, sustained by 18 additional scholars to direct the journal's outputs. SBE reveals one individual as “Editors-in-Chief” assisted by a group of 28 memberships designated as “editor” and 120 among the “Editorial Review Board”. For being a designation, which could be applied to all members of a Board, to a medium group of editors or only to a single editor, it is possible to deduce it has different connotations and thus, heterogenous responsibilities.

Shared editors, inter and intra-journals

Since editors are prestigious researchers, the more prominent one editor is, the higher the recognition level and more invitations will get for further responsibilities. Within our sample, we searched for scholars taking the editorial job for multiple journals. Among the memberships from our study set, 47 were assigned to editors already on the same Board, i.e. the name appeared more than once in EB and they became responsible for more than one duty in the internal editing process. In order to understand the outlets assigning one, two and three concomitant duties within the same top tier the same personality, Figure 1 was developed.

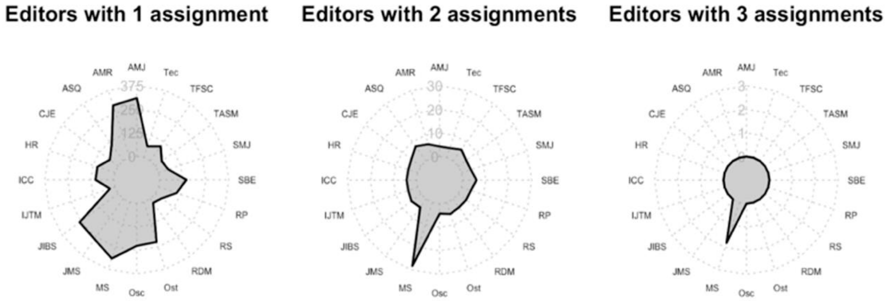


Figure 1. Number of editors in each country, per number of duties.

With exception of MS, all outlets have less than ten editors with two simultaneous assignments. MS has also two editors responsible for three simultaneous assignments in the same Board. Some editors were also found common to more than board. For this journals’ set, the highest number of simultaneous Board memberships held by one individual, is five; thirteen academics hold four Board memberships, 82 have three and 348 have two. From 361 editors in the UK and 1.131 in the US, there are more than 60 and almost 180 editors assuming two duties, respectively. Other studies have also found five as a common number of Board memberships individuals accept simultaneously. Brinn and Jones (2008), in the accounting field, identified two individuals assuming editorial duties for six journals simultaneously and Chan and Fok (2003) found scholar with eight as the maximum number of memberships, in international business.

Clustering of editors based on Boards coupling

In order to address the degree of EB overlapping, we applied some SNA techniques to study the cross-presence of editors within Boards. Based on the so-called ‘interlocking editorship’ phenomenon described by Baccini and Barabesi (2010), the editorial proximity was measured. When a scholar is found in two different Boards, then the two Boards are ‘interlocked’. The interlock bridges the two journals and allows social interaction and communication.

Journals are linked by 627 connections and the density of the interlocking editorship network (i.e. the ratio of the actual number of lines to the maximum possible number of lines in the network) is 41.6%. This is quite superior to the trend previously determined by Baccini and Barabesi (2010) for economic journals and Baccini, Barabesi, & Marcheselli, (2009) for statistical journals.

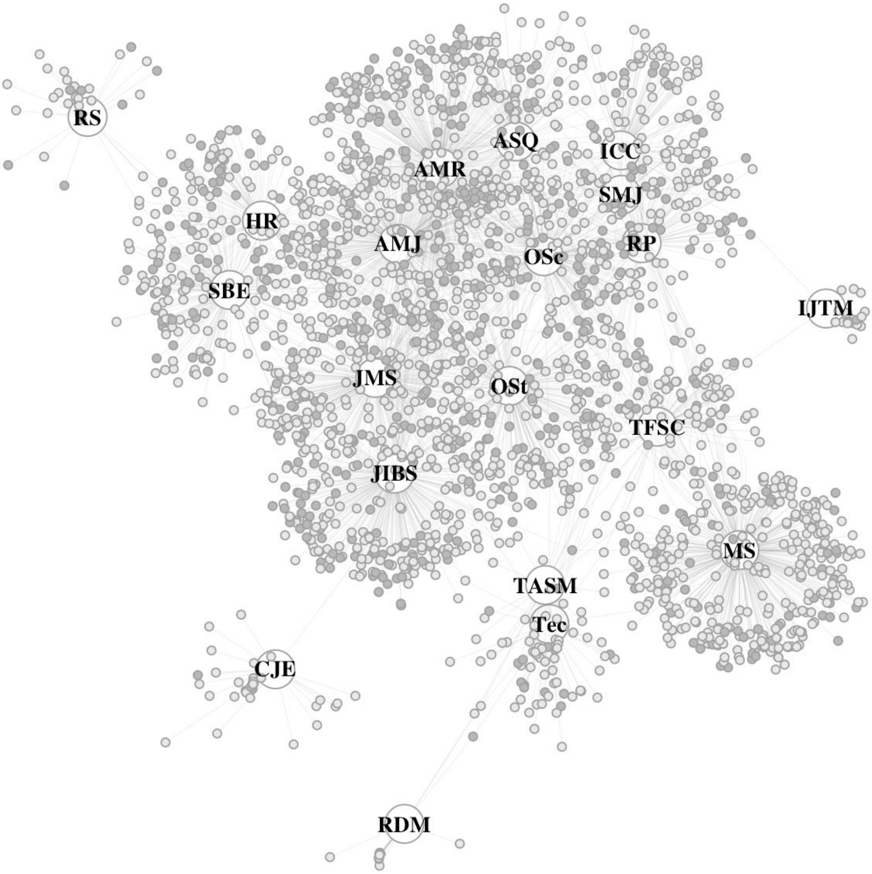


Figure 2. Social network of EB members of innovation-oriented journals.

Using the data from the twenty journals previously identified, we constructed the affiliation network database ad hoc. The average number of seats per journal was 150.25, while the average number of seats occupied by each scholar (i.e., the mean rate of participation) was 1.22. We also investigated the female presence on Boards. The graph of the network is reported in Figure 2, where editors are connected to the journals they work for. Distinct scholars are represented by small nodes (light grey for women and dark grey for man) and their memberships are represented as edges to the big white nodes, the top-tiers.
No journals were found to be completely independent from the others as all outlets are connected, suggesting a strongly connected network. It was also possible to perceive one main group, a giant central which shows close ties formed by editors shared between journal Boards. It is possible to see the high number of scholars holding editorial positions in MS, JIBS, HR and SBE.

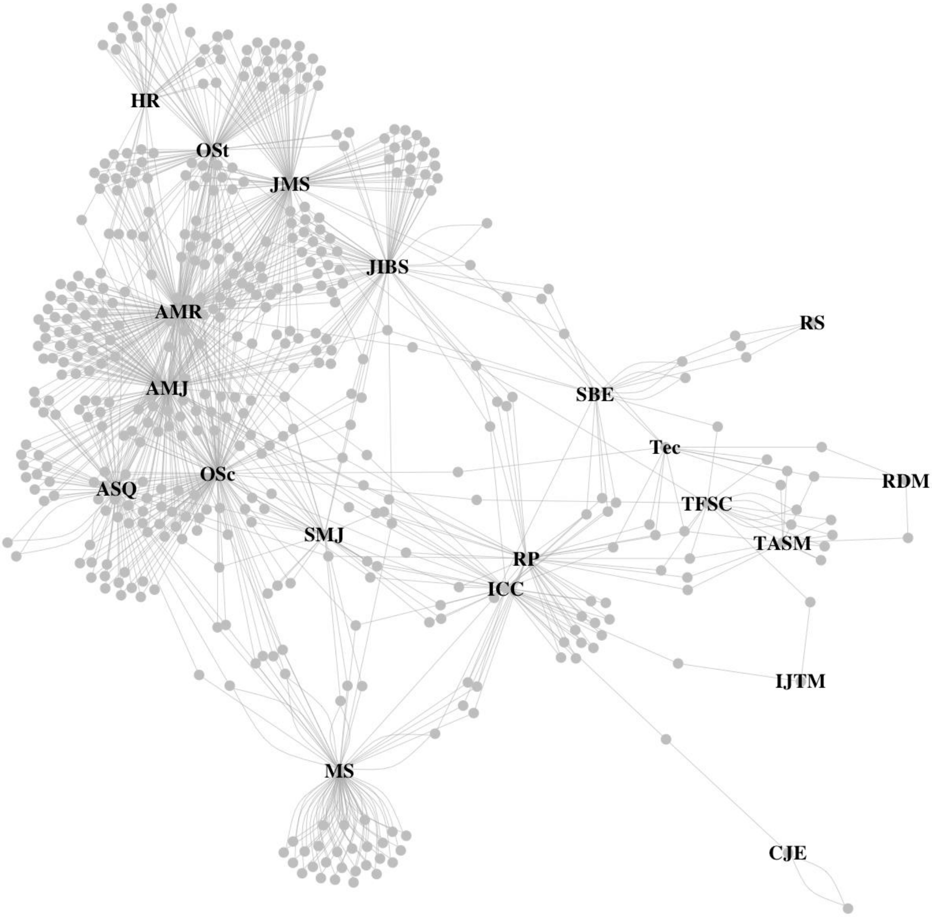


Figure 3. Network illustrating Boards highly connected having staff in common.

More isolated are those journals with lower numbers of editors shared with other journals from the sample: RDM, CJE, RS, IJTM. Actually, CJE shares only one scholar, only with JIBS. Among those more isolated, it is also noticeable the differences on Boards' size. These four outlets have a small number of editors compared to the ones in the central group. Among this network, 73% are men (n=1.783) while 27% are women (n=657). Most female editors are presented in the giant centred sample. Outlets like JIBS, ASQ and SBE show a large number of dark spots surrounded, illustrating the female underrepresentation on those ones. Considering only the editors shared between journals, we plotted Figure 3 to illustrate which outlets share the most scholars.

It is possible to see pairs of journals sharing more editors with larger numbers of edges or larger width edges. Pairs of journals like OSt and MS, ASQ and OSc, RP and ICC share a great number of editors. Regarding the most isolated ones, five journals can be identified: RS, RDM, TASM, IJTM and CJE. The last one, CJE, shares only one editor with ICC. In MS, ASQ and CJE's Boards, it is possible to find some nodes with multiple edges to the same outlet illustrating the multiple editorial roles scholars are assigned to.

The power structures in the interlocking editorship network

One main purpose in SNA is to distinguish between the most central from the peripherical components of a network. In our case, the goal is to perceive which journals are in a central position from those in the boundaries. Centrality analysis may reveal the power and status of the individual or organisation in the social network. As suggested by Wasserman and Faust (1994), three centrality measures may be used: Degree, Closeness and Betweenness.

The simplest measure for the centrality of a journal is represented by the degree of overlap among Boards. Degree centrality is the number of direct connections (lines) a director has with the other journals and measures network influence. It proxies individuals' ability to access, share knowledge or other resources and thereby influence the wider network. Thus, the more ties a journal has to other journals, the more central will be its position in the network.

The Closeness centrality is based on the distance between a journal and all the other journals. This measure calculates the shortest paths between all nodes. A journal is central if its Board can quickly interact with all the other Boards. The more direct and indirect connections a journal has with others, the more central it will be in the network. Journals occupying a central location are best placed to quickly influence the entire network.

Finally, the idea behind the Betweenness is that similar editorial aims between two nonadjacent journals might depend on other journals in the network, especially on those outlets lying on the paths between the two. The number of times a node lies on the shortest path between other nodes is measured as Betweenness centrality. It highlights nodes acting as 'bridges' in a network, proxying for a director's ability to control information and resource flow and to coordinate otherwise disparate parts of the network. In Table 2, centrality measures are provided for all the outlets in the railway network to identify significant top-tiers.

Within our network, we realised both RP and OSc are the outlets with the highest Degree, i.e. with more connections. Considering the normalised degree, which is obtained by dividing the number of connections by the maximum possible number of journals, we realise there are twelve top-tiers showing a normalised degree above 1. In this network encompassing twenty top-tiers, the maximum number of journals an outlet could be linked to is 19. Those twelve exceeding 1 reveal they share more than one editor with some other Boards. Figure 4 provides a graphical representation of the network according to journals' degree, betweenness and closeness. Journals with the average shortest distance (closeness) to all other outlets are represented with greater nodes. Grey scale is applied to illustrate betweenness centrality measure where darker colours represent top-tiers linking higher number of other non-adjacent outlet pairs. Edges between journals are larger for journals with higher number of common editors which justifies the degree score.

Table 2. Journals' centrality measures.

Journal	Degree	Degree normalized	Betweenness	Betweenness normalized	Closeness	Closeness normalized
AMJ	26	1.368	11.35	0.033	0.037	0.704
AMR	26	1.368	11.35	0.033	0.037	0.704
ASQ	20	1.053	2.65	0.008	0.030	0.576
CJE	4	0.211	0	0	0.022	0.422
HR	14	0.737	0	0	0.027	0.514
ICC	24	1.263	61.02	0.178	0.037	0.704
IJTM	6	0.316	0.80	0.002	0.024	0.463
JIBS	24	1.263	9.58	0.028	0.037	0.704
JMS	22	1.158	12.85	0.038	0.033	0.633
MS	20	1.053	3.01	0.009	0.033	0.633
OSc	28	1.474	21.71	0.063	0.038	0.731
OSt	20	1.053	6.09	0.018	0.031	0.594
RDM	6	0.316	0	0	0.021	0.396
RS	4	0.211	0	0	0.021	0.396
RP	28	1.474	36.71	0.107	0.040	0.760
SBE	20	1.053	40.81	0.119	0.033	0.633
SMJ	20	1.053	2.40	0.007	0.032	0.613
TASM	10	0.526	6.28	0.018	0.027	0.514
TFSC	16	0.842	20.72	0.061	0.031	0.594
Tec	18	0.947	36.69	0.107	0.032	0.613

Through the closeness centrality, we understand how long it will take to spread information from a given node. For being the ones with the highest closeness scores and shortest average distance to other nodes, RP and OSc are the outlets represented by larger nodes. Occupying such a position may suggest they are a reference for other outlets. The smallest nodes found are those representing RS and RDM. ICC is the one bridging the bridge for the highest number of other pairs of journals. Its central position may be explained by its interdisciplinary nature, we mean by the presence of many influential editors who enlarge the number of different links with other top-tiers. ICC seems to have an important role facilitating the communication between innovation-oriented journals as it presents the biggest size on its node. For being the most isolated ones, CJE, IJTM, RDM, RS and TASM have a null betweenness as they cannot be a path for other outlets to interact.

Considering the edges, it is possible to understand the degree scores journals obtained. RP and OSc, the ones with higher score, have the largest number of connections. RP has several edges to other outlets and the one with ICC with a larger width. OSc has also multiple edges connecting other top-tiers, three of them representing a great number of shared editors. The most isolated ones are CJE, IJTM, RDM and RS sharing only one or two editors with other outlets from the network. Interesting to remark is that no journals are isolated or completely apart from the network (i.e. they do not present a zero degree).

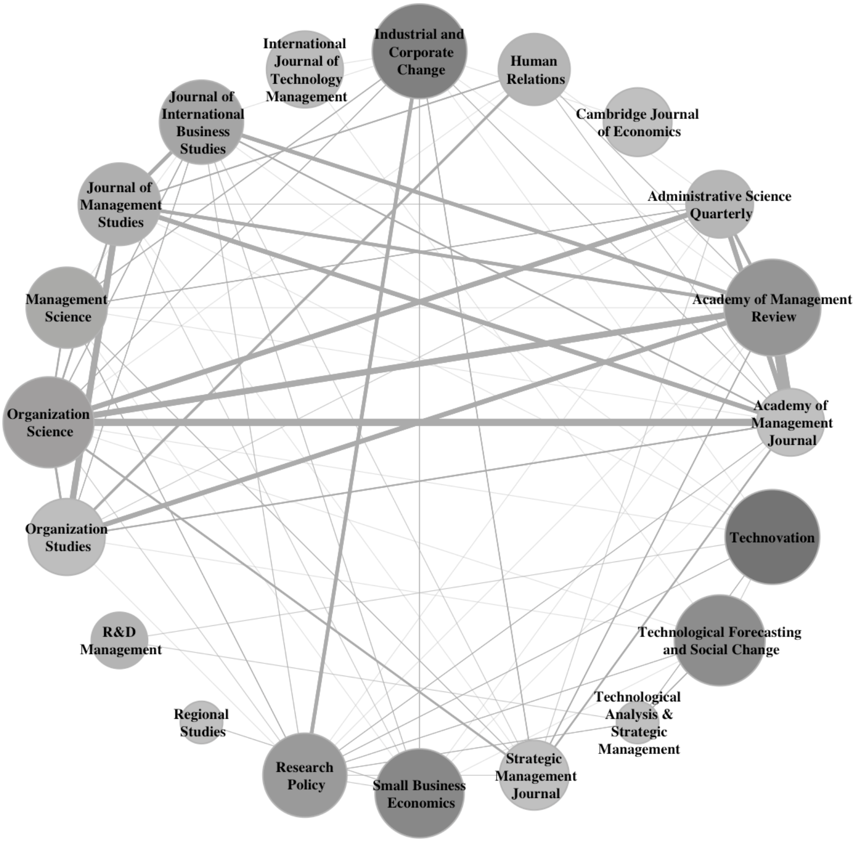

International Journal of Technology Management
Industrial and Corporate Change
Human Relations
Journal of International Business Studies
Cambridge Journal of Economics
Journal of Management Studies
Administrative Science Quarterly
Management Science
Academy of Management Review
Organization Science
Academy of Management Journal
Technovation
Organization Studies
R&D Management
Technological Forecasting and Social Change
Technological Analysis & Strategic Management
Regional Studies
Research Policy
Small Business Economics
Strategic Management Journal

Figure 4. Projection of journals closeness and betweenness centrality and shared editors.

Discussion

In this paper, an overview of editors behind the twenty-top innovation-oriented journals identified by Fagerberg et al. (2012) was provided. We tried to analyse which scientific outlets had the greatest number of editors, which ones share the highest number of scholars with other journals and which top-tiers are the most central ones.

With reference to the internal Board's organisation, the different assignments labels were determined and the heterogeneity in the numbers became manifest. This proliferation of titles without settings description makes it harder to compare the proportion of scholars responsible for the same duties among Boards and responsibilities accepted by an editor present in two Boards. Apart from the Editor-in-Chief role, which may be the single title with the same definition, all the others may entail several different duties and levels of knowledge.

Even though editors have some power shaping the editorial policies and thus, journals sharing editors may have common interests. By measuring the degree of overlap among Boards, the editorial proximity was compared. As an example, AMJ and AMR share 65 Board members, so greater affinity may be expected on their articles published. On the other hand, journals like RDM and CJE have no common editors and no further similar interests are predictable. Actually, by the articles published, we realise the first one is dedicated to Management while the other one to Economics.

Our special interest was to investigate the social relationships between EB members using network techniques. Based on the public data available on journals webpages, we draw our sample of editors from the top twenty innovation-oriented journals and establish co-editorship links between these scholars. Applying a social network methodology, it was possible to provide a rare insight about the dependency of journals through their editors. All outlets were connected at least with another, sharing at least one scholar as editor. This method of research field mapping using co-editorship allowed us to compute centrality measures and provide many novel insights about the relationships between journals' EBs.

By measuring the average geodesic distance, which is the average distance between pairs of nodes in the network, we discovered RP plays a central position for having on Board twenty-eight editors shared with other journals (degree). This journal as well as OSc are able to reach swift communications with other Boards. In addition, RP also revealed to be the closest Board for all the others (closeness). We also determined some journals act as a bridge between others. ICC was the journal occupying the most strategic position facilitating the communication between other pairs of outlets in this twenty journals network (betweenness), followed by the SBE, RP and TFSC. These outlets play a pivotal role connecting top-tiers which do not share editors between their boards and thus, we may infer, with unlike publish policies. By bonding those more distinct channels, these outlets are bringing closer other heterogeneous components which otherwise would be out or disregarded from such a group.

A few limitations should be acknowledged: the official list of EB members on journals' web pages might not be the most updated since a time lag is common between the time a member enters or exits the Board and the appearance of the information in a journal's masthead. Regarding the editorial duties, more detailed analysis of different functions within EB were not possible because of the diversity of positions and the inconsistency of their distributions among journals. Some editorial designations used in one Board are not used in the others, so common assignments are not matchable.

Conclusions

This work analysed the social structure of EB membership in innovation-oriented research based on twenty leading innovation journals in 2019, previously identified by Fagerberg et al. (2012) from 1989 to 2013. The network generated highlighted the presence of shared editors who are responsible for getting journals closer without any independent outlet among this sample.

Regarding the duty's scholars are responsible for, the lack of formalism defining duties allow each journal to decide how to label them. Thus, comparison is not possible. In addition, we considered the number of editors without considering the effort each one of them dedicate to the editorial job.

We can also discuss potential uses of such an approach for science policy purposes and academic governance in our global science system. Co-editorship networks seem prone to reflect intellectual influence of current gatekeepers rather than those who have made significant past contributions but are no longer affiliated to those journals. Comparisons with past Boards' composition may bring more details about outlets' common views and journals connections. Key advantages of social networks encompass the chance to map knowledge wardens in interdisciplinary fields. It could be also used for research fields where literature outputs are published in non-English languages or to map intellectual influence around issues like government policies and scientific processes involving inputs from non-scientific stakeholders but lay experts and others.

In the innovation field, a promising issue to address in the future relates to knowing deeper about those important gatekeepers shared between journals (i.e. where they work or which journals they work with) and determining how common editorial Board composition has put closer editorial policies and similar outputs.

References

Andrikopoulos, A., & Economou, L. (2015). Editorial board interlocks in financial economics. *International Review of Financial Analysis*, *37*, 51–62. Elsevier Inc.

Baccini, A., & Barabesi, L. (2010). Interlocking editorship. A network analysis of the links between economic journals. *Scientometrics*, *82*(2), 365–389. Kluwer Academic Publishers.

Baccini, A., Barabesi, L., & Marcheselli, M. (2009). How are statistical journals linked? A network analysis. *CHANCE*, *22*(3), 35–45. Informa UK Limited.

Bakker, P., & Rigter, H. (1985). Editors of medical journals: Who and from where. *Scientometrics*, *7*(1–2), 11–22. Kluwer Academic Publishers.

Bedeian, A. G., Van Fleet, D. D., & Hyman, H. H. (2009). Scientific Achievement and Editorial Board Membership. *Organizational Research Methods*, *12*(2), 211–238. SAGE PublicationsSage CA: Los Angeles, CA.

Brinn, T., & Jones, M. J. (2008). The composition of editorial boards in accounting: A UK perspective. *Accounting, Auditing & Accountability Journal*, *21*(1), 5–35. Emerald Group Publishing Limited.

Burgess, T. F., & Shaw, N. E. (2010). Editorial Board Membership of Management and Business Journals: A Social Network Analysis Study of the Financial Times 40. *British Journal of Management*, *21*(3), 627–648. John Wiley & Sons, Ltd.

Burns, T., & Stalker, G. (1961). *The Management of Innovation*. London: Tavistock Publications.

Cancino, C. A., Merigó, J. M., & Coronado, F. C. (2017). A bibliometric analysis of leading universities in innovation research. *Journal of Innovation & Knowledge*, *2*(3), 106–124. Elsevier.

Chan, K. C., & Fok, R. C. W. (2003). Membership on editorial boards and finance department rankings. *Journal of Financial Research*, *26*(3), 405–420. Wiley Subscription Services, Inc., A Wiley Company.

de Andrade, R. L., & Rêgo, L. C. (2018). The use of nodes attributes in social network analysis with an application to an international trade network. *Physica A: Statistical Mechanics and its Applications*, *491*, 249–270. North-Holland.

Fagerberg, J., Fosaas, M., & Sapprasert, K. (2012). Innovation: Exploring the knowledge base. *Research Policy*, *41*(7), 1132–1153. North-Holland.

Fagerberg, J., Martin, B., & Andersen, E. (2013). *Innovation Studies: Evolution and Future Challenges* (1st ed.). Oxford: Oxford University Press.

Fagerberg, J., Mowery, D., & Nelson, R. (2004). Innovation: A Guide to the Literature. *The Oxford Handbook of Innovation* (pp. 1–26). Oxford: Oxford University Press.

Feldman, D. C. (2008). Building and Maintaining a Strong Editorial Board and Cadre of Ad Hoc Reviewers. *Opening the Black Box of Editorship* (pp. 68–74). Palgrave Macmillan UK.

Martin, B. (2012). The evolution of science policy and innovation studies. *Research Policy*, *41*(7), 1219–1239. Elsevier B.V.

Morton, M. J., & Sonnad, S. S. (2007). Women on professional society and journal editorial boards. *Journal of the National Medical Association*, *99*(7), 764–771.

Nelson, R. R. (1959). The Simple Economics of Basic Scientific Research. *Journal of Political Economy*, *67*(3), 297–306. University of Chicago Press.

Rakas, M., & Hain, D. S. (2019). The state of innovation system research: What happens beneath the surface? *Research Policy*, *48*(9), 103787. Elsevier B.V.

Rogers, E. M. (1983). *Diffusion of Innovations* (3rd ed.).

Teixeira, E. K., & Oliveira, M. (2018). Editorial board interlocking in knowledge management and intellectual capital research field. *Scientometrics*, *117*(3), 1853–1869. Springer Netherlands.

Topaz, C. M., & Sen, S. (2016). Gender Representation on Journal Editorial Boards in the Mathematical Sciences. *PLoS ONE*, *11*(8). Public Library of Science.

Vespignani, A. (2018). Twenty years of network science. *Nature*, *558*(7711), 528–529. Nature Publishing Group.

Wasserman, S., & Faust, K. (1994). Social Network Data: Collection and Applications. *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Wilkes, M. S., & Kravitz, R. L. (1995). Policies, practices, and attitudes of North American medical journal editors. *Journal of General Internal Medicine*, *10*(8), 443–450

Journals' agendas versus actual publications: A first look at article dynamics in innovation journals

Ana Teresa Santos[1] and Sandro Mendonça[2]

[1] *atmss@iscte-iul.pt*
Instituto Universitário de Lisboa (ISCTE-IUL) – Business Research Unit (BRU-IUL), Lisbon, Portugal.

[2] *sfm@iscte-iul.pt*
Instituto Universitário de Lisboa (ISCTE-IUL) – Business Research Unit (BRU-IUL), Lisbon, Portugal.
UECE/REM – ISEG/ University of Lisbon, Lisbon, Portugal.
SPRU, University of Sussex, UK.

Abstract

In this paper, we address the problem faced by researchers attempting to decide the appropriate journal to submit their works for. Based on content analysis, we studied how semantically similar are journals blurb's sections with the articles published by the outlets. By considering such a methodology, we propose a new strategy for journal's selection for manuscript submission decision based on endogenous outcomes instead of traditional ones like journal's scores centered on dissemination achievements. Throughout, we illustrate our analysis with data from twenty current innovation-oriented journals. We use the articles published from 2010 to 2019 to develop a framework for understanding how historical contents shape publication opportunities for researchers. We emphasize the usefulness of contents already published to understand journals' selection practices. Current statistical approaches to content analysis can grasp the usefulness of already published abstract articles or journal blurbs section as a path to drive further submission decisions and offer reliable measures of influence that may have potential policy implications.

Introduction

Turning a scientific manuscript into a published article and reaching multiple stakeholders is the greatest desire of any scientist. However, it is a tough decision to select an appropriate outlet to submit a research work. In the modern world where both knowledge and technology progress promptly, delays in the publication or wrong audience envisioned might negatively impact a yearly academic review and prevent a pioneering idea from entering the desired field.

By their side, journals publish articles selected by editors who ultimately depend on reviewers' opinions. Expert reviewers evaluate the rigor and value of new discoveries to gauge how they advance the field. Such peer-review constitutes an important approach to evaluating scientific output and it will continue to play a critical role in many forms of evaluation. For having an inbuilt quality filter, journal articles seem to be the most appropriate unit to count.

However, peer review is limited by its subjective nature and weakly correlates with the manuscripts true value (Starbuck, 2005). As a result, highly prestigious journals are publishing a considerable number of low-value articles while lower prestige ones are distributing some admirable papers. This random editorial selection process is also making outstanding manuscripts receive sequential rejections from different journals before being accepted.

The first attempts to describe the motivations of authors reassemble to the 1950s and 1960s when De Solla Price (1963) treating science as a measurable entity, developed some quantitative techniques and introduced the scientometrics concept. Later and to realise the main interests of authors when selecting a journal for submission purposes, Kochen and Tagliacozzo (1974) identified five basic factors which intervene in the choice of a journal: relevance, acceptance rate, circulation, prestige, and publication lag. Within the years, many other studies contributed for the corpus of knowledge in multiple perspectives. For instances, it was perceived publication timelines are field-dependent. Björk and Solomon (2013) determined submission-to-publication times were approximately twice as long in business and economy as in chemistry and the same happened also for earth sciences and chemistry (Garg, 2016).

The number of factors driving this judgement also changed as well as their importance. Rowlands and Nicholas (2005) found the top factors for senior authors were the journal's reputation, readership, and impact factor (IF). Solomon and Björk (2012) also surveyed authors to evaluate the importance of different factors when considering a journal. The most important factor was whether the research fit the scope of the journal, followed by the journal's quality/impact, speed of review and time-to-publication, type of readership, open access option, and likelihood of acceptance. Salinas and Munch (2015) reported journal prestige, likelihood of acceptance, turnaround time, target audience and IF.

Along with those studies, multiple web services were developed to support authors selecting a publishing venue. Besides of being free to use, they support scholars offering multiple measurements of performance to be used as filters for journals selection. Cofactor Journal Selector, created by the London-based firm Cofactor, leads users through a detailed list of filters to match author's publishing requirements (*Journal Selector | Edanz Group*, n.d.). Others like Elsevier Journal Finder or EndNote Match developed by Elsevier and Thomson Reuters respectively, requires user to input key pieces of information about the article to publish (e.g. title, abstract, keywords/phrases) and uses it to find the best matching journals (Kang et al., 2015). However, services provided by publishers are limited to their own pool of publications, assuming authors begin their decision process by first picking a publisher (Forrester et al., 2017).

To the best of our knowledge, scholarly journals are currently compared through four different means: 1) based on directly available information like IF calculated yearly by the Institute of Scientific Information; 2) data from publishers including acceptance rates, number of subscribers and Web load statistics; 3) data calculated from openly available information such as publishing fees and mean time from submission to publication; and 4) data obtained via surveys with authors who have experienced in publishing in a specific journal (Björk & Öörni, 2009).

Since all these methods are not content sensitive, we propose to match descriptive data ('external') with the 'real fine content' ('internal') of a set of journals based on semantic document classification strategies. Taking advantage of the short promotional statement self-prepared by the journals, the blurbs, as well as the journals' portfolio made available through the acknowledged sources of scientific information like the Web of Science platform, we aim we propose a novel recommendation system for those seeking to publish their scientific work.

Bringing some content-sensitivity to journals' comparison

The exponential growth of the number of scientific publications accompanied with the huge progression on the number of scientific outlets makes it difficult for researchers to decide about the journal to choose for publishing their works (Bornmann & Mutz, 2015; Evans, 2013; Gu & Blackmore, 2016; Shiffrin et al., 2018). Several metrics were developed to measure journals impact, importance and ultimately to guide science makers about the channels to submit their novels (Bornmann & Marx, 2015).

Journal IF, proposed by Garfield (1972), was one important metrics guiding authors (Eugene Garfield, 2006). In ecological field, 85,6% of authors revealed that a journal with a high IF is a 'very important' to 'important' criterion when selecting an outlet to submit their works (Aarssen et al., 2008). And, in a large-scale survey, covering 923 scientific journals between 2006 and 2008, it was found a resubmission pattern suggesting a flow from higher to lower IF journals (Calcagno et al., 2012). Seeking to maximize citation counts, researchers choose to submit their works to the journal with the highest IF and then work down the IF list as the manuscript is rejected. However, this kind of strategy ignores the value of following actions and the opportunity costs involved (i.e., every time a paper is rejected). It is seen as a poor predictor of the ultimate success (Wang et al., 2013). Among the IF's limitations, it should be considered as a measure of scientific use by other researchers rather than scientific quality (Callaway, 2016).

However, one of the limitations noticed was the content insensitivity. Disregarding the matter of which journals are made of, it biases the authors' decisions about submissions. Computer-aided screening and analysis might help to both overview and classify contents more efficiently. Already applied in multiple environments ranging from tweets and other personal opinions to sports news and movie reviews, machine learning algorithms has enabled efficient text classification encompassing subjective sentence contents, source identification and sentiment expression classification.

Machine learning in general allows for classifying data into specifically predefined classes based on manually annotated training data. The algorithm then can identify by itself key features specific of the defined classes. For text analysis, the words used, and their frequencies commonly represent the entry data. This method is called "bag of words". The current work was planned to assess the current potential of machine learning to extract content from articles published and compare it with journals blurb section. Machine learning encompasses several models that are implemented in code in different ways. For this purpose, we selected the Rain Forest model.

Innovation-oriented journals as the interest topic

More than half a century old, innovation studies as a research field emerged from different disciplines as Economics (Nelson, 1959), Management (Burns, 1961) and Sociology (Rogers, 1983) bringing such heterogeneity along (Fagerberg et al., 2012). This interdisciplinary was its nature as an eclectic borrowing of cognitive resources was used in its progress (Martin, 2012). Journals from multiple scientific fields are publishing articles related to innovation studies topic.

With the purpose of identify the outlets which accept and publish the most works on such an area of interest, Fagerberg et al. (2012) studied both the most productive as well as those using further innovation findings. Supporting the establishment of innovation studies as a field, we restricted our analyses to the twenty most influential journals Fagerberg et al. (2012) identified. Table 1 catalogues the principal set of outlets with their subject area, year of launch and quartile they occupy currently according to the IF achieved.

Although almost all journals are in the first quartile (the exception is *Technology Analysis & Strategic Management*) evidencing a prestigious position among their pairs, the subject categories where they are coming from vary greatly. The Business, Management and Accounting category attracts sixteen of our outlets and Economics, Econometrics and Finance are the common interest for three. Both *Administrative Science Quarterly* and *Human Relations* fit on Arts and Humanities and *Regional Studies* are set on Environmental Science and Social Sciences categories. For single journals, fields like Computer Science, Psychology and Engineering are also targets.

Believing such a set of outlets has sufficient in common to be addressed together but also varied interests to be distinguishable, we propose to use their both advertisement section, named *blurb*, and produced contents as objects for publication contents comparison. The current project aims to evaluate the potential of machine learning to assess the similarity between the contents published and the ones described in the blurb section designed to attract authors interest for submission.

Table 1. List of journals.

Short name	Journal	Subject area
AMJ	Academy of Management Journal	Business, Management and Accounting
AMR	Academy of Management Review	Business, Management and Accounting
ASQ	Administrative Science Quarterly	Arts and Humanities; Social Sciences
CJE	Cambridge Journal of Economics	Economics, Econometrics and Finance
HR	Human Relations	Arts and Humanities; Business, Management and Accounting; Social Sciences
ICC	Industrial and Corporate Change	Economics, Econometrics and Finance
IJTM	International Journal of Technology Management	Business, Management and Accounting; Computer Science; Engineering; Social Sciences
JIBS	Journal of International Business Studies	Business, Management and Accounting; Economics, Econometrics and Finance
JMS	Journal of Management Studies	Business, Management and Accounting
MS	Management Science	Business, Management and Accounting; Decision Sciences
OSc	Organization Science	Business, Management and Accounting
OSt	Organization Studies	Business, Management and Accounting
RDM	R&D Management	Business, Management and Accounting
RS	Regional Studies	Environmental Science; Social Sciences
RP	Research Policy	Business, Management and Accounting; Decision Sciences; Engineering
SBE	Small Business Economics	Business, Management and Accounting; Economics, Econometrics and Finance
SMJ	Strategic Management Journal	Business, Management and Accounting
TASM	Technology Analysis & Strategic Management	Business, Management and Accounting; Decision Sciences
TFSC	Technological Forecasting and Social Change	Business, Management and Accounting; Psychology
Tec	Technovation	Business, Management and Accounting; Engineering

Developing a model for blurbs classification

Model conception

This section presents the details of the proposed techniques for matching abstracts from articles published with blurbs sections. The rationales behind these techniques selected are also discussed. Since the accuracy of machine learning algorithms is known to improve with greater quantities of data to train on, articles' abstracts were used to train the model and test it for overall accuracy determination. Later, the algorithm will be applied to match journals blurbs to the related articles abstracts.

In this work, we propose to use a supervised text classification method as we can train the model with the articles' abstracts extracted associated to the outlets titles which published them. The algorithm will learn from the labelled training data to predict outcomes for unforeseen data.

The full process is illustrated in Figure 1 and encompasses five smaller steps which will be addressed in sequence.

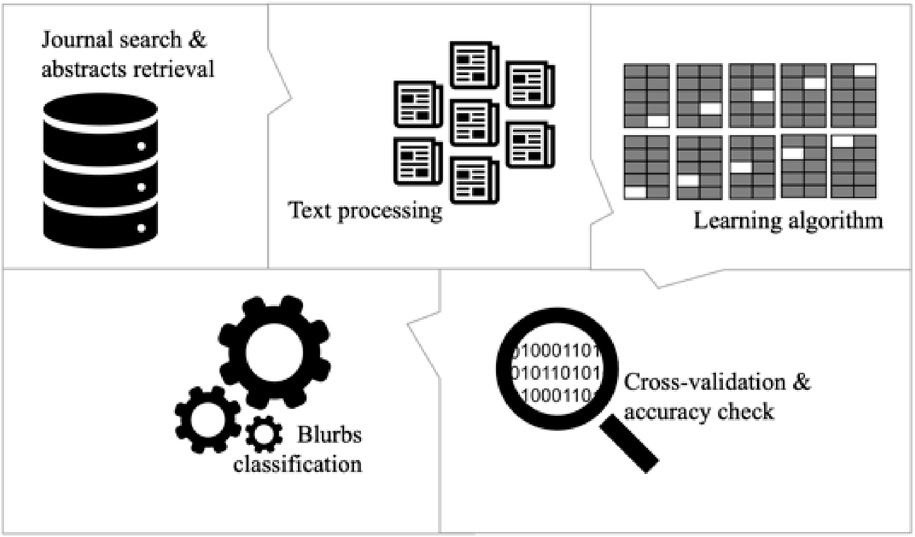


Figure 1. Supervised learning process applied.

With this approach, a training abstracts dataset (previously classified into journals) is used to identify unique patterns that represent each top-tier, and then use these identified patterns to correctly predict the outlet a future instance will belong to.

Journal's search & abstracts retrieval

We searched on Web of Science for the all the titles published in these twenty outlets from 2010 to 2019 (Norris & Oppenheim, 2007). For each one, we extracted the title, the list of authors, the abstract and year of publishing. Only articles with all these contents available were considered. At the end, we realised each journal published different numbers of articles per year and it changed over the years. The Figure 2 shows the number of articles published in each outlet during this ten-years period and the overall volume of articles published for all of them.

Outlets presented different number of published articles and an overall trend to increase over the years. Both TFSC and MS presented a significant increase in the number of accepted manuscripts for publication, publishing less than 150 articles in 2010 and publishing more than 300 in 2019. Some titles were also found with a more modest growth such as AMR and CJE which presented 28 and 64 articles published in 2010 and 34 and 70 in 2019, respectively.
Such increasing trend was anticipated by de Solla Price (1961) who first published quantitative data about the growth of science from 1650 to 1950 with a growth rate of 5.6% per year and forecasted an increasing in journals number reaching one million by 2000. Also interesting are outlets like IJTM which published more in 2010 than in 2019.

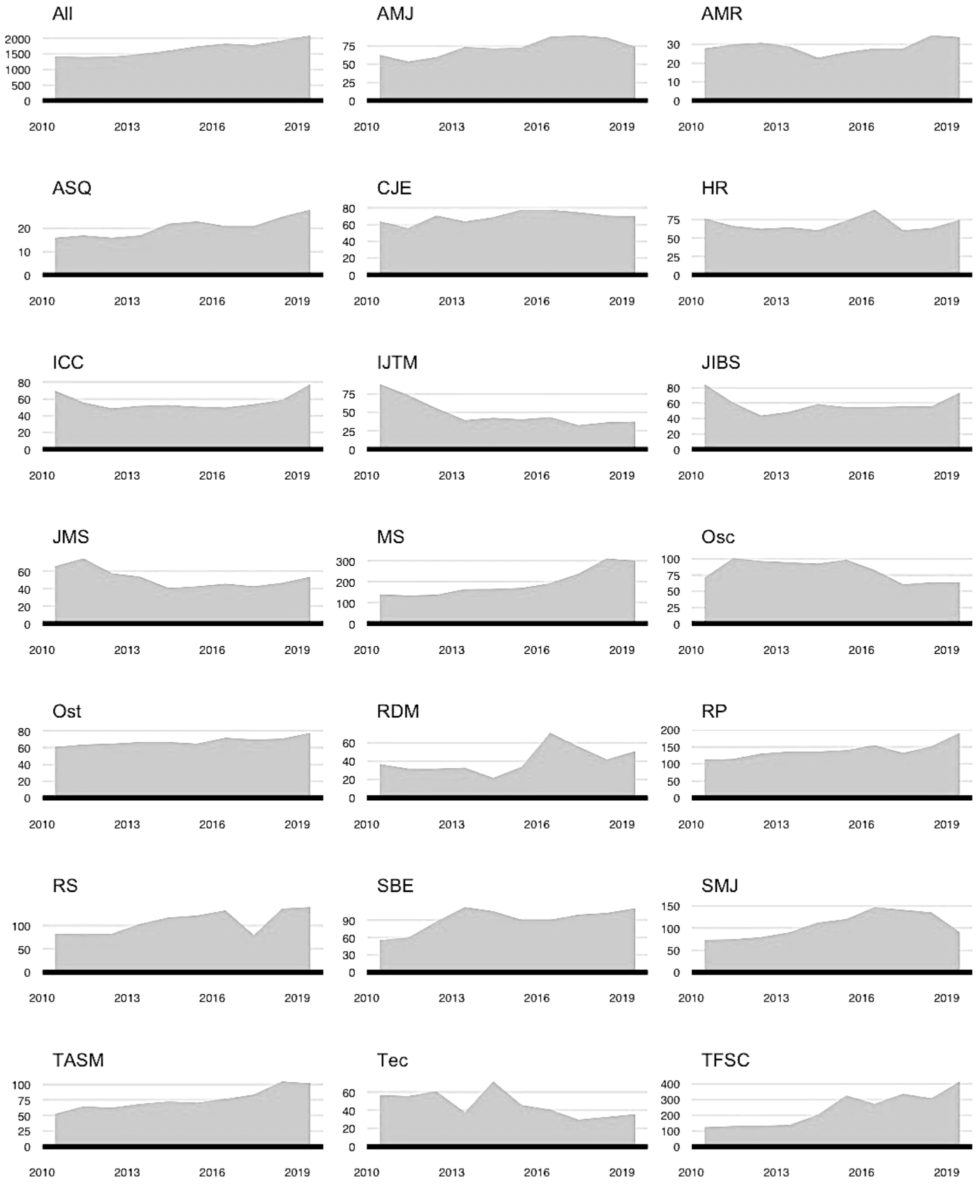


Figure 2. Number of articles published in each outlet from 2010 to 2019.

Text processing

As the classifiers and learning algorithms cannot directly process the text documents in their original form, we need to transform our raw text documents with variable length into numerical feature vectors with a fixed size. Therefore, abstracts collected were pre-processed and converted into a more manageable representation.

This first major step involved four smaller tasks: 1) word tokenization; 2) case transformation; 3) stop-word removal and 4) stemming. Word Tokenization aims to separate words as tokens.

In a simple way to perform tokenization is to assume [space] as separators between words or tokens. Case transformation involves changing all capitalized letters in a lower-case format. Stop-word removal aims to delete the most frequent words that occur in the English language like *the*, *and*, *a*, *is*... In addition, we also removed "paper" as it was quite frequent. These so common words do not bring any meaning to the text which allow us to exclude them without undesired impact. Finally, stemming reduces the inflected words to their root form or stem. In this study, we applied the Porter Stemmer algorithm, one of the most used stemming algorithms (Porter, 1980).

The pre-process step is considered a critical one as it affects the speed and accuracy of a learning algorithm. At the end, it is expected datasets no longer have high levels of noise or unmeaning stems which will be used as features. With this clean dataset, each feature (word) was mapped to the corresponding number of occurrences (term-frequency) within the whole text. This is called "word embedding" process and converts each abstract to a vector of the same length containing the frequency of the words. These vectors will be used to train the model and later, classify blurbs.

Learning algorithm

Among the supervised learning algorithms known, those combining multiple learning models using either boosting or bagging techniques have shown to achieve better results than simple learning models (Caruana & Niculescu-Mizil, 2006). For this work, Random Forests was the one which outperform other learning algorithms (Breiman, 2001). Although it does not consider the sequence of instances to be classified in sequential labelling tasks, it does not present a major problem for our project.

While correcting for decision trees' inherent problem of over-fitting of training examples, Random Forest was operated by constructing a multitude of decision trees during training, then outputting the class that was the mean prediction of the individual decision trees. Like any other supervised learning techniques, the goal of Random Forest model is to identify patterns in a training set and then use these identified patterns to predict future unseen cases.

The main assumption of machine learning is that the distribution of training data is identical to the distribution of test data and future examples. If the learning algorithm accurately classifies the set used for testing, then the machine learning assumption suggests that it will perform as well for future unseen cases. For the training purposes, our classification dataset made of abstracts was split into two disjoint sets and 70% of the observations were used for training and 30% for testing.

Cross-validation & accuracy check

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

Training accuracy was determined by a ten-fold cross validation (Arlot & Celisse, 2010; Kohavi, 1995). In a ten-fold cross validation, the input data is divided into ten equal disjoint subsets. Each subset is used as the test set while all the others are used as the training set. For each validation, accuracy is measured by the ratio between the "number of correctly classified cases" and the "total number of cases". The final accuracy will be the average accuracy of the ten different subsets. Depending on the application the learning model is designed for, the minimum acceptable accuracy level may be different. The overall accuracy of our model was 80% which means our model can correctly predict four out of five abstracts. Previous articles have described and used accuracy measures that are widely recognized as the standard way to determine the test and training accuracies of prediction models (Mullen & Collier, 2004).

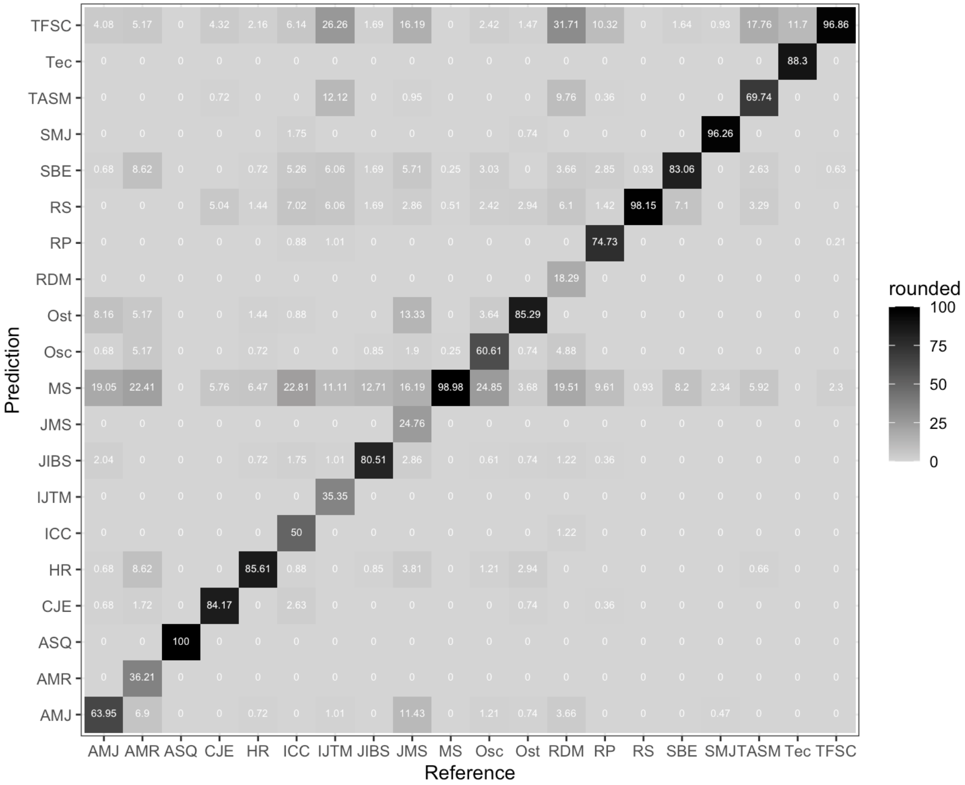


Figure 3. Abstracts' classification.

Although the overall accuracy was 80%, the accuracy achieved by model classifying the abstracts belonging to each specific journal varies greatly. Figure 3 shows all the *Administrative Science Quarterly*'s abstracts are correctly classified. Closer to this figure are the abstracts published on *Management Science* (98,98%), *Regional Studies* (98,15%) and *Technological Forecasting and Social Change* (96,86%). Nevertheless, the model performance was not so good predicting the abstracts published on *R&D Management*, *Journal of Management Studies*, *International Journal of Technology Management* and *Academy of Management Review*.
Such accuracy differences presented by the model are explained by the type of contents published. The more specialized journals are greater the predictable accuracy achieved by the model. Journals publishing a larger number of heterogeneous topics are classified as the publishers of abstracts from other outlets. This happens with *Management Science* and *Technological Forecasting and Social Change*. Both are predicted to have published great numbers of abstracts from other journals. Stating their eclectic and diverse interest, their editorial announcement is not only being defined as generalist but also as direct competitors of other journals.

Blurbs' classification

Our analysis showed that the current model can be efficiently used to match abstracts with the journals which published them. Abstracts previously collected were used to train and test the model, the algorithm was trained again considering this larger dataset. All pre-process steps were repeated before the training. This time the test dataset was the blurbs sections collected from the outlets. The model was asked to associate the blurbs according to the most semantically similar journal abstracts collection. Figure 4 presents the classifications suggested by the model.

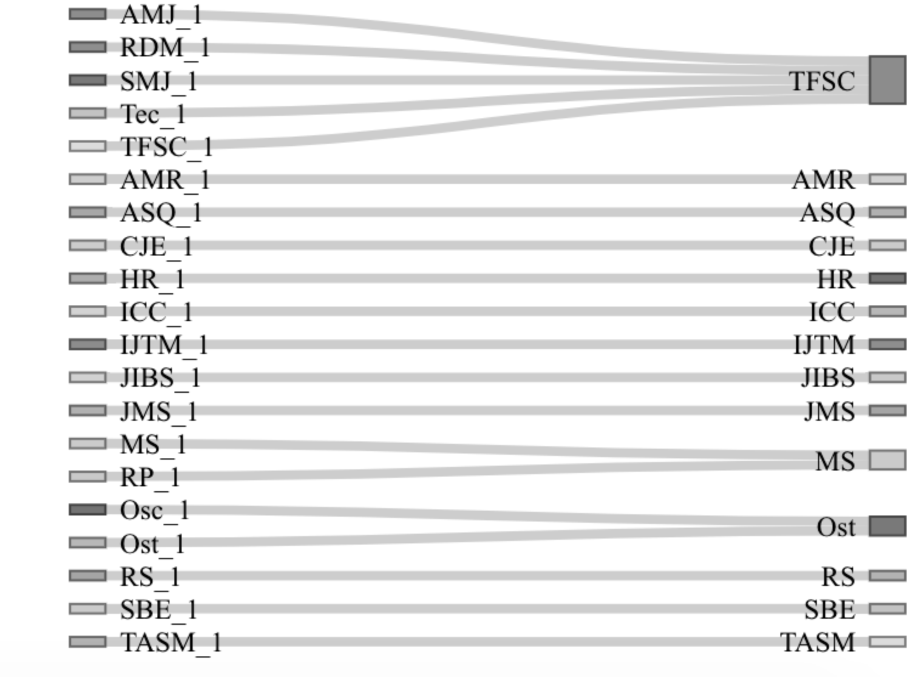


Figure 4. Blurbs' classification.

Within the twenty journals sample, fourteen blurbs were correctly linked to the journals they belong to. *Technological Forecasting and Social Change, Management Science* and *Organization Studies* were associated with blurbs from other journals besides their own's. The model recognized *Management Science* and *Research Policy* blurbs sections like *Management Science's* abstracts from ten years of published articles. The model also recognized blurbs from *Academy of Management Journal*, *R&D Management*, *Strategic Management Journal*, *Technovation* and *Technological Forecasting and Social Change* as being related to the last one. Blurbs from *Organization Science* and *Organization Studies* were also associated with *Organization Studies* published abstracts.

Following such results, an author may also wonder about submit a manuscript designed for *Research Policy* also to *Management Science*. It seems manuscripts from both top tiers are poorly differentiated and confusing to distinguish the outlets which published them. Thus, all manuscripts are classified as part of *Management Science* portfolio. As both are in Q1 according to the IF in their fields, a manuscript in the scientific area of *Research Policy* may be also interesting for *Management Science.* The same could be also true for the five journals which had their blurbs' section associated with the contents of *Technological Forecasting and Social Change.* Considering the contents previously published besides the journals' external metrics available may provide additional opportunities for researcher submissions.

Discussion

Given the constantly growing stream of scientific journals' data, automatic classification of unstructured text data into relevant researcher-defined content categories is likely to continue charming scientists. The results of our analysis both confirms it is possible to recognize journals contents by their blurbs and suggests blurbs are distinctive pieces of information able to judge journals interests.

In terms of the accuracy of determining the web page type using machine learning techniques, Random Forest achieved up to 80% accuracy. The automatic classifier, however, misclassifies articles from *Management Science* journal. The number of abstracts used to train the model cannot explain this behaviour as this journal was pointed out as one the journals with higher number of articles published over the years and with a significant increase. Our suggestion is that *Management Science* publishes a great variety of scientific topics misleading the algorithm when recognising the journals which published a manuscript.

Organization Studies showed to be also the outlet which could also publish the works a research may think about submit to *Organization Science. Technological Forecasting and Social Change* seems to be the outlet which could encompass choices from four different top-tiers: *Academy of Management Journal*, *R&D Management*, *Strategic Management Journal*, *Technovation* and *Technological Forecasting and Social Change*. The sematic similarity between these ones, make *Technological Forecasting and Social Change* an alternative option to the three others.

There are of course limitations to this research. First, we considered only the ten years of data from articles available on Web of Science from the Innovations Studies, a field which emerged a few decades ago. Considering other older scientific fields and larger data published may bring more accurate classifications. Second, for the algorithm training, we used only the abstracts of full articles which may differ from the usage of full articles contents. All these issues are also ideas for further studies to enlarge even more the scientometric approaches able to provide contents sensitive methods for journals comparison.

Concluding Remarks

Text classification, that is computer-aided analysis of textual data, offers a great opportunity to advance journal selection strategies. This motivated us to apply such technique to a real problem: compare published contents with the advertisement section designed to attract further submissions.

However, some limitations should be reported and overcome in the future works. First and foremost, the content period of this project is set as ten years (ranging between 2010 and 2019), which was mainly driven by the consideration of balancing the size of data and the computational capacity.

We applied a machine learning due to its great potential for efficient classification. However, a drawback was noticed for this method. It lacks transparency, which prevents identification of causal factors (Liu et al., 2019).

For future exploration of the potential of machine learning algorithms to compare journal contents, a larger set of annotated training abstracts, blurbs and journals would be necessary. For the current proof of concept, the "Abstracts" and Blurbs sections were manually taken from journals' websites. Automatic extraction will efficiently generate a larger data sets and may bring such methodology to other areas of science knowledge, benefiting researchers from areas beyond innovation studies.

References

Aarssen, L. W., Tregenza, T., Budden, A. E., Lortie, C. J., Koricheva, J., & Leimu, R. (2008). Bang for Your Buck: Rejection Rates and Impact Factors in Ecological Journals. *The Open Ecology Journal*, *1*(1), 14–19. https://doi.org/10.2174/1874213000801010014

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Björk, B. C., & Öörni, A. (2009). A Method for Comparing Scholarly Journals as Service Providers to Authors. *Serials Review*, *35*, 62–69. https://doi.org/10.1080/00987913.2009.10765213

Björk, B. C., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, *7*(4), 914–923. https://doi.org/10.1016/j.joi.2013.09.001

Bornmann, L., & Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics*, *9*(2), 408–418. https://doi.org/10.1016/j.joi.2015.01.006

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222. https://doi.org/10.1002/asi.23329

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burns, T. (1961). *The management of innovation*. Tavistock Publications.

Calcagno, V., Demoinet, E., Gollner, K., Guidi, L., Ruths, D., & de Mazancourt, C. (2012). Flows of research manuscripts among scientific journals reveal hidden submission patterns. *Science*, *338*(6110), 1065–1069. https://doi.org/10.1126/science.1227833

Callaway, E. (2016). Beat it, impact factor! Publishing elite turns against controversial metric. *Nature*, *535*(7611), 210–211. https://doi.org/10.1038/nature.2016.20224

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *ACM International Conference Proceeding Series*, *148*, 161–168. https://doi.org/10.1145/1143844.1143865

de Solla Price, D. J. (1961). *Science since Babylon*. Yale U.P.

de Solla Price, D. J. (1965). *Littel Science, Big Science*. Columbia University Press.

Evans, J. A. (2013). Future science. *Science*, *342*(6154), 44–45. https://doi.org/10.1126/science.1245218

Fagerberg, J., Fosaas, M., & Sapprasert, K. (2012). Innovation: Exploring the knowledge base. *Research Policy*, *41*(7), 1132–1153. https://doi.org/10.1016/J.RESPOL.2012.03.008

Forrester, A., Björk, B.-C., & Tenopir, C. (2017). New web services that help authors choose journals. *Learned Publishing*, *30*(4), 281–287. https://doi.org/10.1002/leap.1112

Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science*, *178*(4060), 471–479. https://doi.org/10.1126/science.178.4060.471

Garfield, Eugene. (2006). The History and Meaning of the Journal Impact Factor. *JAMA*, *295*(1), 90. https://doi.org/10.1001/jama.295.1.90

Garg, K. C. (2016). Publication delay of manuscripts in periodicals published by CSIR-NISCAIR. *Current Science*, *111*(12).

Gu, X., & Blackmore, K. L. (2016). Recent trends in academic journal growth. *Scientometrics*, *108*(2), 693–716. https://doi.org/10.1007/s11192-016-1985-3

Journal Selector | Edanz Group. (n.d.). Retrieved July 12, 2019, from https://www.edanzediting.com/journal-selector

Kang, N., Doornenbal, M., & Schijvenaars, B. (2015). Elsevier journal finder: Recommending journals for your paper. *RecSys 2015 - Proceedings of the 9th ACM Conference on Recommender Systems*, 261–264. https://doi.org/10.1145/2792838.2799663

Kochen, M., & Tagliacozzo, R. (1974). Matching authors and readers of scientific papers. *Information Storage and Retrieval*, *10*(5–6), 197–210. https://doi.org/10.1016/0020-0271(74)90059-X

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*, *14*(12), 1137–1143.

Liu, Y., Chen, P. H. C., Krause, J., & Peng, L. (2019). How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA*, *322*(18), 1806–1816. https://doi.org/10.1001/jama.2019.16489

Martin, B. (2012). The evolution of science policy and innovation studies. *Research Policy*, *41*(7), 1219–1239. https://doi.org/10.1016/J.RESPOL.2012.03.012

Mullen, T., & Collier, N. (2004). Incorporating topic information into sentiment analysis models. *Empirical Methods in Natural Language Processing*, 412–418.

Nelson, R. R. (1959). The Simple Economics of Basic Scientific Research. *Journal of Political Economy*, *67*(3), 297–306. https://doi.org/10.1086/258177

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics*, *1*(2), 161–169. https://doi.org/10.1016/j.joi.2006.12.001

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130– 137. https://doi.org/10.1108/00330330610681286

Rogers, E. M. (1983). *Diffusion of Innovations* (3rd ed.).

Rowlands, I., & Nicholas, D. (2005). Scholarly communication in the digital environment: The 2005 survey of journal author behaviour and attitudes. *Aslib Proceedings*, *57*(6), 481–497. https://doi.org/10.1108/00012530510634226

Salinas, S., & Munch, S. B. (2015). Where Should I Send It? Optimizing the Submission Decision Process. *Plos One*, *10*(1), e0115451.

Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2632–2639. https://doi.org/10.1073/pnas.1711786114

Solomon, D. J., & Björk, B.-C. (2012). Publication fees in open access publishing: Sources of funding and factors influencing choice of journal. *Journal of the American Society for Information Science and Technology*, *63*(1), 98–107. https://doi.org/10.1002/asi.21660

Starbuck, W. H. (2005). How much better are the most-prestigious journals? The statistics of academic publication. *Organization Science*, *16*(2), 180–200. https://doi.org/10.1287/orsc.1040.0107

Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132. https://doi.org/10.1126/science.1237825

Quantum technology 2.0 – topics and contributing countries from 1980 to 2018

Thomas Scheidsteger [1,] Robin Haunschild[1], Lutz Bornmann[2] and Christoph Ettl[2]

[1] *{t.scheidsteger,r.haunschild}@fkf.mpg.de*
Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70659 Stuttgart (Germany)

[2] *{lutz.bornmann,christoph.ettl}@gv.mpg.det*
Administrative Headquarters of the Max Planck Society, Hofgartenstr.8, 80539 Munich (Germany)

Abstract

The second quantum technological revolution started around 1980 with the control of single quantum particles and their interaction on an individual basis. These experimental achievements enabled physicists and engineers to utilize long-known quantum features - especially superposition and entanglement of single quantum states - for a whole range of practical applications. We use a publication set of 54,598 papers from the Web of Science published between 1980 and 2018 to investigate the time development of four main subfields of quantum technology in terms of numbers and shares of publication as well as the occurrence of topics and their relation to the 25 top contributing countries. Three successive time periods are distinguished in the analyses by their short doubling times in relation to the whole Web of Science. The periods can be characterized by the publication of pioneering works, the exploration of research topics, and the maturing of quantum technology, respectively. Compared to the US, China has a far over proportional contribution to the worldwide publication output, but not in the segment of highly-cited papers.

Introduction

In the beginning of the 20th century Planck's quantum hypothesis to derive the correct black body radiation (Planck, 1901) and Einstein's explanation of the photoelectric effect (Einstein, 1905) led to a full-grown quantum theory in the mathematical formulations of the matrix mechanics of Heisenberg, Born, and Jordan (Born, Heisenberg, & Jordan, 1926) as well as of Schrödinger's wave mechanics. Quantum theory turned out to be highly consistent with experiment. The theory formed the basis for the development of solid state physics and for a first quantum technological revolution. This development led to applications such as lasers, transistors, nuclear power plants, solar cells, and superconducting magnets in NMR devices or particle accelerators. These applications have in common the exploitation of quantum behavior of great ensembles of particles.

In the late 1970s and early 1980s, scientists learned to prepare and to control systems of single quantum particles such as atoms, electrons, and photons, and to let the particles interact on an individual basis. This ability sparked a second quantum revolution, where physicists and engineers worked together to utilize the long-known quantum features - especially superposition and entanglement of single quantum states - for a whole range of practical "next generation" applications. These applications may be summarized as "quantum engineering" or "quantum technology 2.0" (QT 2.0).

The present study provides a bibliometric analysis of QT 2.0 methodologically following previous studies which have dealt with research fields such as climate change in general (Haunschild, Bornmann, & Marx, 2016), specific aspects thereof (Marx, Haunschild, & Bornmann, 2017a; Marx, Haunschild, & Bornmann, 2017b, 2018), and density functional theory (Haunschild, Barth, & Marx, 2016). This study analyzes QT 2.0 over the time period 1980-2018 with a focus on the following four topical fields:

(i) *Quantum information science* (Q INFO): The proposition by Einstein, Podolsky, and Rosen (1935) that quantum systems can exhibit non-local, entangled correlations unknown in the classical world could be experimentally proven after 45 years (Aspect, Grangier, & Roger, 1982; Clauser & Shimony, 1978). Since then it is the basis of quantum information processing, e.g., by the use of non-local photon correlations to make an unbreakable quantum cryptographic key distribution over a long optical fibre (Tapster, Rarity, & Owens, 1994). The concept of a quantum bit or qubit (the quantum mechanical generalization of a classical bit) can be physically realized as a two-state device, exploiting the coherent superposition of both states. The engineering challenge is the layout of hardware systems that are able to handle many qubits in quantum gates, registers and circuits. In particular, the qubits must be stored and kept stable enough to perform several computation cycles in order to realize a quantum computer.

(ii) Quantum metrology and sensing (Q METR): New measurement techniques provide higher precision than the same measurement performed in a classical framework. One example is a new generation of quantum logic clocks achieving a previously unknown accuracy by exploiting the sensitivity of quantum entanglement against disturbances. Other examples are atom interferometry-based gravimeters (Snadden, McGuirk, Bouyer, Haritos, & Kasevich, 1998) and magnetic field sensors based on quantum defects in diamonds (Barry et al., 2016). Control of quantum systems is achieved via manipulating quantum interferences of the wave functions in coherent laser beams. The control is guided by the so-called quantum optimal control theory (Brif, Chakrabarti, & Rabitz, 2010).

(iii) Quantum communication and cryptography (Q COMM): The field started by the publication of the BB84 protocol for quantum key exchange by Bennett and Brassard (1984). BB84's main component is the quantum key distribution via entangled qubits which would render under cover eavesdropping impossible. Quantum networks consist of quantum processors which exchange qubits over quantum communication channels. Secure communication in quantum networks is essential for the long-range transmission of quantum information usually by quantum teleportation.

(iv) Quantum computing (Q COMP): This field promises a quantum leap in computational power, since previous speed-ups as described by Moore's law (Moore, 1995) appear to come to an end on the basis of semiconductor technology (Bilal, Ahmed, & Kakkar, 2018). The original idea of quantum computing had been expressed by Feynman (1982): quantum systems as, e.g., molecules should be simulated by letting a model quantum system evolve and calculate the system in question. That was a new approach – rather different from implementing the classical algorithms of, e.g., quantum chemistry, which consume a high amount of computational resources. The first implementation of quantum simulation had been the quantum variant of simulated annealing, a widely used Monte Carlo optimization. In 2011, the Canadian enterprise D-Wave announced to have built the first commercial quantum annealer (Johnson et al., 2011). Others try to implement a universal model of quantum computation using quantum logic gates in superconducting electronic circuits. They tried to reach "quantum supremacy" by means of demonstrating that a programmable quantum device can solve a problem that no classical computer can feasibly solve. In 2019, Google has claimed to have reached this goal (Arute et al., 2019) with its quantum processor for a very special problem, for which the world's largest supercomputer would take thousands of years. New algorithms and software are necessary to exploit the advantages of quantum computing.

Tolcheev (2018) published a bibliometric study on quantum technology including a very broad set of papers (by comprising all papers that use "quantum" in their title, abstract, or keywords). In contrast, this study has a more focused view by including papers from specific technology-relevant subfields. This study does not consider many foundational quantum mechanics/physics applications, quantum chemistry applications, and other quantum related concepts that are too unspecific. These applications and concepts would decrease the precision of the search.

Methods and Data set

Data set

The bibliometric data used in this study are from three sources: (1) the online version of the Web of Science (WoS) database provided by Clarivate Analytics (Philadelphia, Pennsylvania, USA), (2) the bibliometric in-house database of the Max Planck Society (MPG), developed and maintained in cooperation with the Max Planck Digital Library (MPDL, Munich), and (3) the bibliometric in-house database of the Competence Centre for Bibliometrics (CCB, see: http://www.bibliometrie.info/). Both in-house databases were derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI) , Conference Proceedings Citation Index-Science (CPCI-S), and Conference Proceedings Citation Index-Social Science & Humanities (CPCI-SSH) prepared by Clarivate Analytics. The analyses considered publications of the document types "Article", "Conference Proceeding“, and "Review". The results are based on 54,598 papers published between 1980 and 2018. The papers had been searched in the online version of the WoS on 25 May 2020 by using 14 subqueries related to QT 2.0. The subqueries have been carefully constructed for sufficient recall and high precision (Bornmann, Haunschild, Scheidsteger, & Ettl, 2019).

Publication output, citations indicators, and mapping of research topics

We analyzed the number of papers (full counting) broken down by year, field of QT 2.0, and country. Citation impact analyses are based on time- and field-normalized indicators. We focused on the share of papers belonging to the 10% most frequently cited papers in the corresponding publication year, document type, and subject area. In case of more than one paper with a citation count at the required threshold of 10%, these papers are assigned fractionally to the top 10% publication set. This procedure ensures that there are exactly 10% top 10% papers in each subject area (Waltman & Schreiber, 2013). The top-10% indicator is a standard field-normalized indicator in bibliometrics (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015). The citation window relates to the period from publication until the end of 2018.

Besides indicators such as publication and citation counts as measures of scientific activity and impact, techniques of text mining are also used in bibliometric studies. The analysis of keywords in a corpus of publications can identify important research topics and reveal their development over time. This analysis can be managed with the software package VOSviewer (van Eck & Waltman, 2010) which produces networks based on bibliographic coupling. The nodes in these networks are keywords, their size signifies the number of corresponding publications, and the distance between nodes is proportional to their relatedness regarding cited references. Keywords of papers citing similar literature are located closer to each other. The nodes are divided into classes of similarity, displayed by clusters of different colors. The network can be controlled by some adjustable parameters such as minimal cluster size or resolution.

Results

Respective share and overall growth of fields

We retrieved 54.598 publications using the search queries. Table 1 shows the number of papers in the four fields, their percentages of the total number of publications (which add up to more than 100% because of overlaps between the fields), and the percentage of papers belonging to only one field.

Table 1. Number and percentage of papers in four fields of QT 2.

Field of QT 2.0	*Number of papers*	*Percentage of the number of distinct papers*	*Number and percentage of one-field-only papers*
i) Q INFO	16,300	29.85%	9,706 (59.55%)
ii) Q METR	12,531	22.95%	9,766 (77.93%)
iii) Q COMM	13,985	25.61%	9,809 (70.14%)
iv) Q COMP	21,786	39.90%	16,545 (75.94%)
Sum of all fields		118.77%	45,826 (83.93%)

Figure 1 shows the annual publication counts for QT 2.0 and its four fields for the period from 1990 to 2018. The annual numbers of publications on QT 2.0 before 1990 never exceeded a dozen per year. Especially papers about Q METR were published before 1990, which can be explained by the efforts and achievements in manipulation, controlling and measuring of single quantum systems. An exponential growth until about 2000 can be seen in Figure 1, mainly caused by Q METR and Q COMP. From 2000 to 2011, Q COMP is the most strongly represented field, with about twice as many papers as Q INFO and Q COMM.

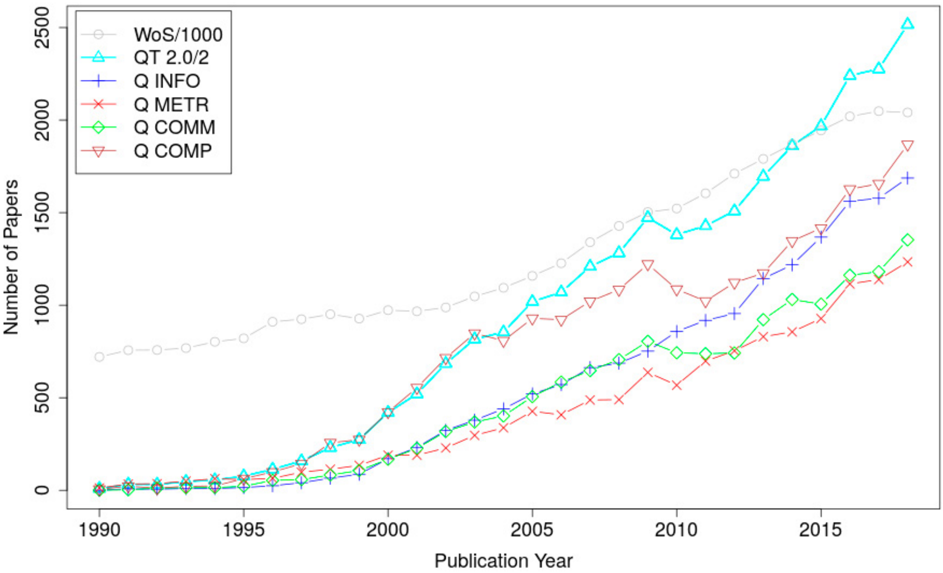


Figure 1: Annual numbers of publications of QT 2.0 and its four fields between 1990 and 2018 (in earlier years the annual numbers of all fields together never exceed 12) compared to the number of articles, reviews, and conference proceedings in the whole WoS. The numbers for QT 2.0 and the whole WoS are scaled by factors of 2 and 1000, respectively, for better comparison.

To study the time evolution, we divide the period in three phases: (1) from 1980 to 1999, (2) from 2000 to 2011, and (3) from 2012 to 2018. The numbers of papers in the last two periods are on the same order of magnitude with 24,322 and 28,132. With 2,144 papers, the first pioneering phase has only less than a 10th of the paper numbers in the other periods. A measure of the growth of the research fields during the three time periods is the doubling time (see Table 2). The four fields have very similar values as the total QT 2.0. The very short doubling time of two years is characteristic for the first period until 1999, slowing down to four years in the second period and to seven years in the most recent period. The last doubling time is comparable to the 5-6 years which Haunschild, Bornmann, et al. (2016) found for the climate change literature until 2014. However, this time is significantly shorter than the 12-13 years for the overall growth of the WoS records. Bornmann and Mutz (2015) calculated an even higher doubling time of nearly 24 years for the WoS in the time period from 1980 to 2012 by applying a non-linear segmented regression analysis. During the 20 years from 1991 to 2010 the annual number of publications grew by a factor of 42, compared to a factor of ten for the climate change corpus and to a factor of about two for the whole WoS.

Table 2: Doubling times in years of the three time periods for all QT 2.0 papers and the four fields compared to the whole WoS publication records.

Time period	*All QT papers*	*Q INFO*	*Q METR*	*Q COMM*	*Q COMP*	*WoS*
1980-1999	2-3	1-2	3-4	1-2	1-2	7-8
1980-2011	4-5	4-5	4-5	4-5	5-6	11-12
1980-2018	6-7	5-6	6-7	6-7	7-8	12-13

Contributing Countries

Many countries are contributing research on QT 2.0 by collaborating with each other. The 25 top publishing countries (with at least 500 papers published between 2000 and 2018 in QT 2.0) give a similar picture in the four subfields and in QT 2.0 as a whole – with nearly the same countries dominating. The same 22 countries are among the top 25 countries in QT 2.0 and *all four* fields, even when we focus on the 10% most cited papers in QT 2.0 and its fields. For both cases (either all papers or only the top 10% papers), we calculated two numbers: the first number is the difference (%QT - %WoS); positive and negative signs mean more or less publication activity than expected. The second number is the corresponding quotient (%QT / %WoS). The quotient is identical to the so-called Activity Index (AI) which was introduced by Frame (1977). AI is a variant of the Revealed Comparative Advantage (RCA) used in economics (Mittermaier et al., 2017). AIs greater than 1.0 indicate national publication outputs higher than expected (from the whole WoS). Both indicators are presented as radar charts in Figure 2 each containing a plot including all papers and the top 10% papers. In each radar chart, the 22 common countries are denoted by their respective two-letter country codes. The codes start at the top with the country having the most publications in QT 2.0 (US) and descend clockwise. In each radar chart, the dividing values between under and over achievement are marked by a grey dashed line at the value 0 for the difference and 1 for the AI.

The most striking insight from these figure is the very different assessment of the two leading countries with very similar output, the US and China, in comparison with the whole WoS: while the US is less active in QT 2.0 than in other WoS-covered research fields (QT 2.0: Difference = -5.7%, AI = 0.77), China is much more active in QT 2.0 (QT 2.0: Difference = +6.9%, AI = 1.71). The difference is most pronounced in Q COMM (Difference = +15.2%, AI = 2.5). With respect to top 10% papers, the strong research focus of China on QT 2.0 is dampened considerably (QT 2.0: Difference = +1.9%, AI = 1.26; Q COMM: Difference = +7.4%, AI = 2.0).

Figure 2 shows that Austria, Singapore, and Switzerland contribute rather high shares of QT 2.0 research in comparison with their research activities as a whole. Austria has an overall AI of above 2 in QT 2.0 and Q COMP, and of nearly 3 in Q COMM and Q INFO. The AIs even exceeded by focusing on the top 10% papers, leading to values of more than 4. These high AIs are explainable by high activities of research groups in Vienna and Innsbruck concerning quantum teleportation. Singapore has AI values of nearly 3 in Q INFO, Q COMM, and Q COMP. Switzerland's AI value of about 1.6 is mainly caused by a high value of 2.4 in Q COMM.

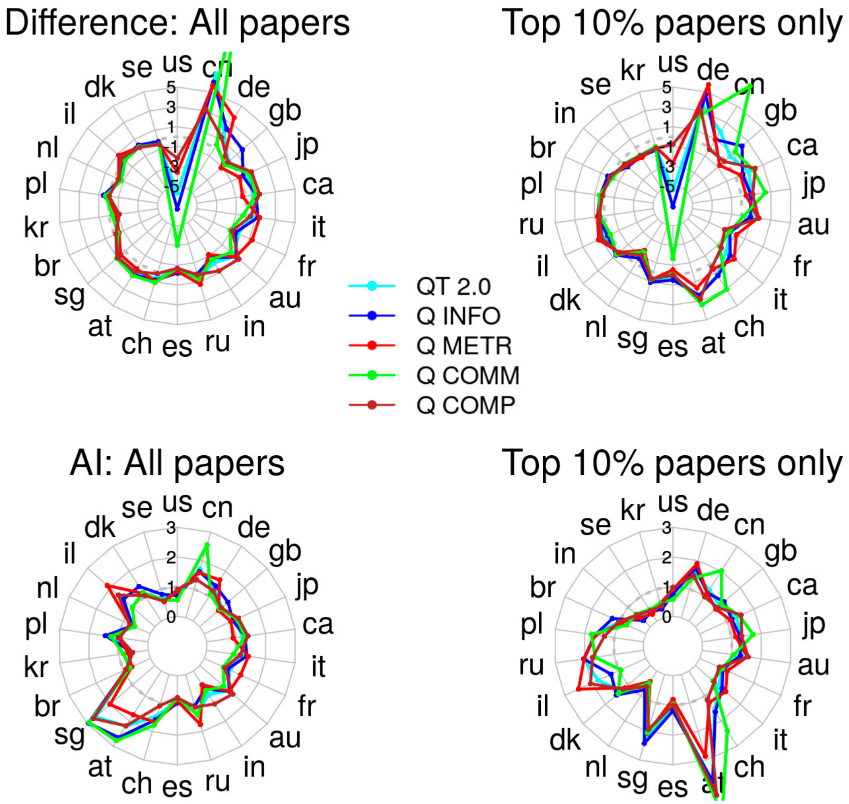


Figure 2: Radar charts of the differences (upper graphs) and quotients (activity indices, lower graphs) of the national shares of papers in QT 2.0 and its four fields (for the period 2000-2018). On the left side, all papers are included; on the right side, only the top 10% most cited papers in the time period 2000-2016 are included. The 22 countries, which are among the top 25 in all fields and the whole QT 2.0, are denoted by their country codes, and ordered clockwise in descending order of the number of publications. The grey dashed lines at 0 and 1 indicate the expected output of the country. An online version can be viewed at https://tinyurl.com/yy75v77a.

We investigated to what extent the growth of QT 2.0 research in specific countries was transferred into the area of applications. This can be measured by the share of publications with commercial co-authorships. Since the databases that we used for this study only provide this information for German affiliations, we focused on the subset of publications with authors only from Germany. Because of the very small number of QT 2.0 publications in the first (pioneering) period until 1999, we only consider the second and third period starting from 2000. The share of publications with German commercial co-authorships is between about 4.5 and 5.5% in the whole WoS with German authorship alone from 2000 to 2018. The QT 2.0 share of these publications during the exploration period from 2000 to 2011 is about 2.3%, but during the maturing period from 2012 to 2018 it is nearly twice as high with 4.3%.

Visualization of research topics and their time evolution

For the various time periods, we have created keyword maps based on author keywords and keywords plus assigned by Clarivate Analytics. A common thesaurus file (https://tinyurl.com/y35eqaax) was used to unify singular/plural forms of words and synonyms. The minimal number of keyword occurrences is chosen such that about 100 keywords are displayed for each time period. The VOSviewer parameters for the clustering have been chosen to be default values. However, for the minimal cluster size, we used a value of 5 to receive a well interpretable network. All VOSviewer maps are provided as Java-based web-startable versions (Oracle, 2020) via URLs for an interactive inspection by the readers, e.g. by zooming into the clusters.

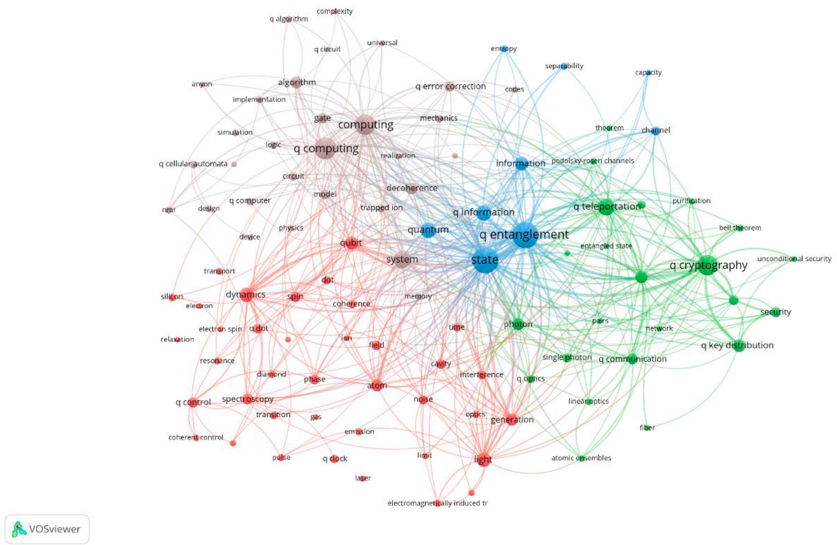


Figure 3: Co-occurrence map of the top 100 keywords (author keywords and keywords plus) in the period 1980 to 2018 with four topical clusters, using the VOSviewer parameters resolution=1.0 and minimal cluster size=5. (The two biggest unnamed green nodes belong to the keywords communication and key distribution.) For better readability of compound keywords, quantum is abbreviated to q (web-startable version at https://tinyurl.com/y8lvjnyt).

Figure 3 shows the overall co-occurrence map of 100 keywords occurring at least 298 times for the period from 1980 to 2018. Maps with about 100 keywords usually are a good compromise between keeping readability of the map and displaying most of the content. In the figure, the four fields of QT 2.0 are nicely discernible by the *keywords* in four clusters, whose colors are kept consistent in all networks: (1) Red (Q METR) with the *manipulation* of single *atoms*, *molecules*, and even (*electron*) *spins* as in *quantum dots*, and the *quantum control* using *light fields* of *coherence* (*lasers*) that lead to the *realization* of single *qubits* and of very high precision *quantum clocks*. (2) Brown (Q COMP) with q*uantum computing* and *computers* that build on *quantum circuits* with *logic gates* realized as *trapped ions, anyons* or in *NMR* devices. On this hardware, *quantum algorithms* have been implemented that are much in need of *quantum error correction*. (3) Blue (Q INFO) with *quantum entanglement* of *states* that is a major subject of *quantum information science*, also investigating *entropy* and *channel capacity* in a generalization of Shannon's information theory. (4) Green (Q COMM) with q*uantum communication* that is built upon the *quantum teleportation* of *pairs* of *entangled states*, often realized by *single photons*, as a basis for *quantum cryptography* and *quantum key distribution*.

When we inspect the topic maps for the three partial time periods, we see continuity and persistence of clusters as well as change of focus and occurrence of keywords: from the first to the second time period only 60 out of 99 keywords in the maps are identical. From 1980 to 1999 (https://tinyurl.com/y7c2nxcc) the focus had been on the preparation, manipulation, and control of single quantum systems at the atomic scale and the pioneering work on building materials, devices, and sensors for quantum metrology. From 2000 to 2011 (https://tinyurl.com/ycwqehyu) the focus had, on the one hand, switched to the advanced design of hardware components for real quantum computers and the development of algorithms utilizing quantum properties. On the other hand, the exploitation of quantum effects like entanglement for secure communication using quantum key distribution had become prominent (favorably utilizing quantum optics of single photons). From the second to the third time period, i.e. 2012 to 2018 (https://tinyurl.com/y7c2nxcc), nearly 80% of the keywords remain the same (78 out of 99 and 101 keywords, respectively). There are only slight changes in the main direction of research, but some keywords moved into the new clusters of Q INFO and quantum optics. For example, memory and storage are located in the clusters Q COMP and Q METR in the second period, and are connected with quantum optics in the third period. This connection is probably because of their importance for optical quantum communication networks. The keyword quantum simulation appears only on the third map in the Q COMP cluster. This coincides with the enlarged efforts to build a quantum simulator as the fulfilment of Feynman's vision of a quantum computer (Harris et al., 2018; Johnson et al., 2011).

Visualization of the geographical distribution of research topics

Figure 4 shows a combination of the approaches taken in the previous two sections. For the period from 1980 to 2018, we have produced a co-occurrence map of countries (denoted by their two-letter country code with a prefixed "@") with at least 400 occurrences (multiple co-authorships of the same country on one paper are counted only once) as well as a map of keywords (author keywords and keywords plus assigned by Clarivate Analytics) with at least 300 occurrences. These thresholds lead to the above mentioned top 25 countries and to 104 keywords, sorted into five topical clusters (by using the VOSviewer parameters resolution=1.1 and minimal cluster size=5). Four clusters in the figure can be assigned to the four fields of QT 2.0. About ten countries are assigned to the clusters Q METR and Q INFO, respectively. Three countries are assigned each to Q COMP and Q COMM. In case of Q COMP, India and Iran are mainly connected to the *design* of *logic gates* and *circuits*.

We compare now the assessment of the countries in the radar charts for all QT 2.0 papers in Figure 2 with their placement and connection in Figure 4: the large node of China in the green cluster (Q COMM) mirrors the dominance of China in this field with respect to the total number of papers and its high AI of nearly 3. Germany with the third highest publication output and the highest values in Q METR in the radar charts is located consequently prominently in the red cluster. Germany has significant contributions to quantum optics, and is connected to some other countries in the blue cluster (Q INFO). Q INFO is the field of Germany's second highest AI.

We would like to emphasize two other countries. These countries have with 2.5% a small share of all QT 2.0 papers, but a high AI of about 3 in Q INFO and Q COMM: Singapore has a high AI of about 3 in Q INFO and Q COMM. In the map, consequently, it can be found in the blue cluster of Q INFO connected with *quantum entanglement* and *information* (and with the UK).

The country is additionally connected with the green cluster of Q COMM and its major player China. Austria's activities, especially in Innsbruck and Vienna, are mirrored by its placement in the blue cluster Q INFO. This cluster is strongly connected to *quantum entanglement.* It is also connected to the green Q COMM keywords *communication* and *pairs* of *photons* (quantum optics, orange cluster).

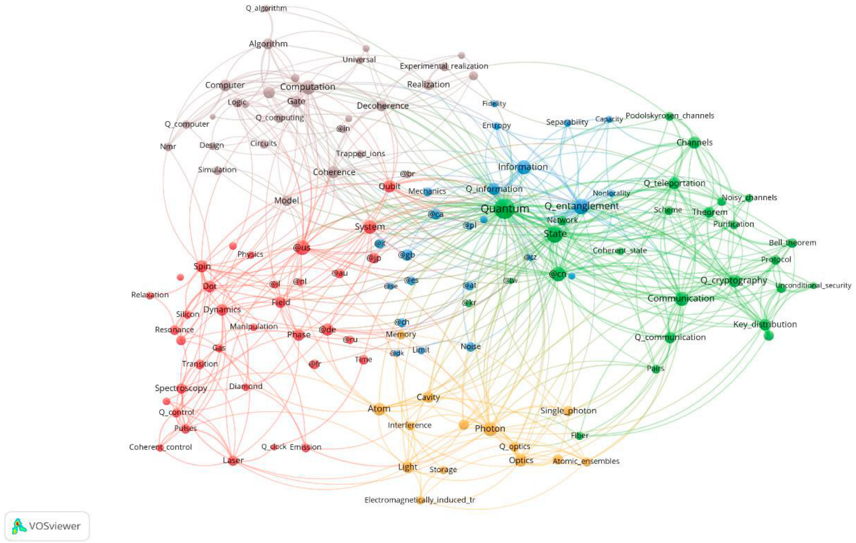


Figure 4: Co-occurrence map of the top 25 countries (denoted by their two-letter country code with a prefixed "@") with at least 400 occurrences and the top 104 keywords (author keywords and keywords plus) with at least 300 occurrences in the total publication set (from 1980 to 2018) with five topical clusters, using the VOSviewer parameters resolution=1.1 and minimal cluster size=5. For better readability of compound keywords, the term Quantum is abbreviated to Q (web-startable version at https://tinyurl.com/y978bpzm).

Discussion and Conclusions

This bibliometric study on QT 2.0 identified four main subject fields, namely Q INFO, Q METR, Q COMM, and Q COMP. For these four fields, we analyzed their respective share and growth. Of the 54.598 publications in our dataset, the four fields have shares from about one fifth (Q METR) to two fifth (Q COMP) (see Table 1). In the first decade considered here, less than 100 publications have appeared dominated by Q METR with its pioneering works on preparing and controlling single quantum systems. During the second decade (the 1990s) Q COMP joined Q METR in driving the exponential growth, leading to the ongoing dominance of Q COMP in the new millennium (see Figure 1). Between 1980 and 1999, the doubling time of QT 2.0 was between 2 and 3 years; the doubling time of the whole WoS is 7 to 8 years. During the time periods until 2011 and until 2018, respectively, the doubling times were about half as long as in the whole WoS, with 4 to 5 years and 6 to 7 years, respectively (see Table 2). In the most recent decade, therefore, QT 2.0 is a very active research area with a steady exponential development, which is common for mature research fields. We also analyzed the main contributing countries to QT 2.0. We focused on a time period with a substantial annual number of papers from 1990 until 2018. We looked at the top 25 contributing countries in more detail and compared their publication output in QT 2.0 and its four fields to the one expected from the whole WoS (see Figure 2). We visualized the geographical distribution of research topics with a co-occurrence map of countries and keywords in Figure 4. The main result is the sharp contrast of the US and China which are the greatest contributors to QT 2.0. The US shows a much smaller contribution to QT 2.0 than could be expected from their otherwise leading role in science. China has a far over proportional contribution, especially in the field of Q COMM – corroborated by its hub-like function in the topical map. By focussing on highly cited publications, China's share and AI are significantly diminished.

In future studies, the transfer of QT 2.0 research into the area of (commercial) applications might be an interesting research question. In this study, we could only investigate this topic for publications with authors exclusively from Germany.

Acknowledgments

The bibliometric data used in this study are from the bibliometric in-house databases of the Max Planck Society (MPG) and the Competence Centre for Bibliometrics (CCB, see: http://www.bibliometrie.info/). The MPG's database is developed and maintained in cooperation with the Max Planck Digital Library (MPDL, Munich); the CCB's database is developed and maintained by the cooperation of various German research organizations. Both databases are derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI) prepared by Clarivate Analytics (Philadelphia, Pennsylvania, USA). We thank one of the reviewers for pointing out the interesting research question concerning commercial applications of QT 2.0.

References

Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J. C., Barends, R., . . . Martinis, J. M. (2019). Quantum supremacy using a programmable superconducting processor. *Nature, 574*(7779), 505-510. doi: 10.1038/s41586-019-1666-5.

Aspect, A., Grangier, P., & Roger, G. (1982). Experimental Realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: A New Violation of Bell's Inequalities. *Physical Review Letters, 49*(2), 91-94. doi: 10.1103/PhysRevLett.49.91.

Barry, J. F., Turner, M. J., Schloss, J. M., Glenn, D. R., Song, Y., Lukin, M. D., . . . Walsworth, R. L. (2016). Optical magnetic detection of single-neuron action potentials using quantum defects in diamond. *Proceedings of the National Academy of Sciences, 113*(49), 14133-14138. doi: 10.1073/pnas.1601513113.

Bennett, C. H., & Brassard, G. (1984, 1984). *Quantum cryptography: Public key distribution and coin tossing*.

Bilal, B., Ahmed, S., & Kakkar, V. (2018). Quantum Dot Cellular Automata: A New Paradigm for Digital Design. *International Journal of Nanoelectronics and Materials, 11*(1), 87-98.

Born, M., Heisenberg, W., & Jordan, P. (1926). Zur Quantenmechanik. II. *Zeitschrift für Physik, 35*(8), 557-615. doi: 10.1007/BF01379806.

Bornmann, L., Haunschild, R., Scheidsteger, T., & Ettl, C. (2019). *Quantum technology – a bibliometric analysis of a maturing research field*.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology, 66*(11), 2215-2222. doi: 10.1002/asi.23329.

Brif, C., Chakrabarti, R., & Rabitz, H. (2010). Control of quantum phenomena: past, present and future. *New Journal of Physics, 12*(7), 075008. doi: 10.1088/1367-2630/12/7/075008.

Clauser, J. F., & Shimony, A. (1978). Bell's theorem. Experimental tests and implications. *Reports on Progress in Physics, 41*(12), 1881–1927. doi: 10.1088/0034-4885/41/12/002.

Einstein, A. (1905). Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Annalen der Physik, 322*(6), 132-148. doi: 10.1002/andp.19053220607.

Einstein, A., Podolsky, B., & Rosen, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review, 47*(10), 777-780. doi: 10.1103/PhysRev.47.777.

Feynman, R. P. (1982). Simulating physics with computers. *International Journal of Theoretical Physics, 21*(6), 467-488. doi: 10.1007/BF02650179.

Frame, J. D. (1977). Mainstream research in Latin America and the Caribbean. *Interciencia, 2*, 143–148.

Harris, R., Sato, Y., Berkley, A. J., Reis, M., Altomare, F., Amin, M. H., . . . Yao, J. (2018). Phase transitions in a programmable quantum spin glass simulator. *Science, 361*(6398), 162-165. doi: 10.1126/science.aat2025.

Haunschild, R., Barth, A., & Marx, W. (2016). Evolution of DFT studies in view of a scientometric perspective. *Journal of Cheminformatics, 8*, 12. doi: 10.1186/s13321-016-0166-y.

Haunschild, R., Bornmann, L., & Marx, W. (2016). Climate Change Research in View of Bibliometrics. *Plos One, 11*(7), 19. doi: 10.1371/journal.pone.0160393.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429-431.

Johnson, M. W., Amin, M. H. S., Gildert, S., Lanting, T., Hamze, F., Dickson, N., . . . Rose, G. (2011). Quantum annealing with manufactured spins. *Nature, 473*(7346), 194-198. doi: 10.1038/nature10012.

Marx, W., Haunschild, R., & Bornmann, L. (2017a). Climate change and viticulture - a quantitative analysis of a highly dynamic research field. *Vitis, 56*(1), 35-43. doi: 10.5073/vitis.2017.56.35-43.

Marx, W., Haunschild, R., & Bornmann, L. (2017b). The Role of Climate in the Collapse of the Maya Civilization: A Bibliometric Analysis of the Scientific Discourse. *Climate, 5*(4). doi: 10.3390/cli5040088.

Marx, W., Haunschild, R., & Bornmann, L. (2018). Climate and the Decline and Fall of the Western Roman Empire: A Bibliometric View on an Interdisciplinary Approach to Answer a Most Classic Historical Question. *Climate, 6*(4). doi: 10.3390/cli6040090.

Mittermaier, B., Holzke, C., Tunger, D., Meier, A., Glänzel, W., Thijs, B., & Chi, P.-S. (2017). *Erfassung und Analyse bibliometrischer Indikatoren für den PFI-Monitoringbericht 2018.*

Moore, G. E. (1995). LITHOGRAPHY AND THE FUTURE OF MOORE LAW. In J. M. Warlaumont (Ed.), *Electron-Beam, X-Ray, Euv, and Ion-Beam Submicrometer Lithographies for Manufacturing V* (Vol. 2437, pp. 2-17). Bellingham: Spie-Int Soc Optical Engineering.

Oracle (2020). Java Web Start Documentation, from https://docs.oracle.com/javase/8/docs/technotes/guides/javaws/ (accessed 12 January 2021)

Planck, M. (1901). Ueber das Gesetz der Energieverteilung im Normalspectrum. *Annalen der Physik, 309*(3), 553-563. doi: 10.1002/andp.19013090310.

Snadden, M. J., McGuirk, J. M., Bouyer, P., Haritos, K. G., & Kasevich, M. A. (1998). Measurement of the Earth's Gravity Gradient with an Atom Interferometer-Based Gravity Gradiometer. *Physical Review Letters, 81*(5), 971-974. doi: 10.1103/PhysRevLett.81.971.

Tapster, P. R., Rarity, J. G., & Owens, P. C. M. (1994). Violation of Bell's Inequality over 4 km of Optical Fiber. *Physical Review Letters, 73*(14), 1923-1926. doi: 10.1103/PhysRevLett.73.1923.

Tolcheev, V. O. (2018). Scientometric Analysis of the Current State and Prospects of the Development of Quantum Technologies. *Automatic Documentation and Mathematical Linguistics, 52*(3), 121-133. doi: 10.3103/S000510551803007X.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538. doi: 10.1007/s11192-009-0146-3.

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology, 64*(2), 372-379. doi: 10.1002/asi.22775.

Does Wikipedia Cover the Relevant Literature on Major Innovations Timely? An Exploratory Case Study of CRISPR/Cas9

Marion Schmidt,[1] Wolfgang Kircheis,[2] Arno Simons,[3] Martin Potthast[2] and Benno Stein[4]

[1]German Center for Higher Education Research and Science Studies (DZHW), schmidt@dzhw.eu
[2]Leipzig University, wolfgang.kircheis@uni-leipzig.de, martin.potthast@uni-leipzig.de
[3]DZHW, Humboldt Universität zu Berlin, arno.simons@posteo.de
[4]Bauhaus-Universität Weimar, benno.stein@uni-weimar.de

Abstract

This *research-in-progress* paper analyzes Wikipedia's representation of the Nobel Prize winning CRISPR/Cas9 technology to explore to what extent and with what temporal dynamics Wikipedia cites the most relevant and visible scientific literature on this topic. We use both verbatim and fuzzy matching heuristics to match publications cited from a selection of secondary formats—like reviews—as well as field-delineated highly cited publications with the central Wikipedia article on CRISPR. Our methodical results confirm that a combination of verbatim searches by title, DOI, and PMID is sufficient. Initial evidence also shows that the Wikipedia article references a substantial amount of articles that are well acknowledged by experts and highly cited, as well as literature that is not strictly scientific or less visible. Delays in coverage on Wikipedia compared to the publication years show a dependence on the dynamics of both the field and the Wikipedia article itself.

Introduction

The Nobel Prize winning development of the CRISPR/Cas9 mechanism and technique, the first steps of which date back more than twenty years, has left a long paper trail of scientific publications and its citation network. To understand outstanding innovation processes such as CRISPR/Cas9 and for science policy to act upon it, the identification of the most relevant publications is of key importance. One way this can be done is by measuring certain properties of the citation network and taking them as proxies for relevance. Another option is to turn to the secondary literature (like reviews) which tries to answer the question of relevance with its specific methodologies, depending on the systematic or narrative format. A problem with the latter is time: Even if a review accurately reflects the relevant literature at the time of its publication, the review may soon become outdated—a problem even more pronounced in the recent development of so-called Living Reviews (Elliot et al., 2014).

In this paper we introduce and explore a third approach to determining the relevant literature of innovation processes like CRISPR/Cas9: the analysis of Wikipedia. The online encyclopedia is a tertiary literature format whose extensive network of articles on scientific subjects is comparable to the concept of a Living Review and hence may be utilized as such for specific domains. Wikipedia is characterized by the fact that it addresses the general public, that articles are constantly revised and updated; thus reflecting dynamics to a greater or lesser extent, as well as by its unique reviewing process. The evolution of references to academic literature on Wikipedia has increased significantly over the years, suggesting that Wikipedia could even be harnessed for altmetrics (Zagovora et al., 2020). Giles (2005), Reavley et al. (2012), Estevez & Cukierman (2012) and Garcia del Valle et al. (2018) suggest that Wikipedia's reflection of scientific knowledge is accurate, while Teplitskiy et al. (2015) as well as Jemielniak et al. (2019) show that top impact journals are most referenced in Wikipedia (medical) articles. For these top biomedical journal papers, it takes (on average) only about three months to be referenced in Wikipedia. Seeing as innovation processes are characterized by transferring scientific knowledge into application contexts with implications on commercial, political, ethical, and legal aspects or questions, connecting wider societal spheres and public perception, these characteristics suggest Wikipedia as an interesting format for representing and mapping innovation processes.

In our paper we use the CRISPR/Cas9 innovation as an in-depth case study to shed light on Wikipedia's referencing patterns. Based on reference corpora matched to the central Wikipedia article on CRISPR we assess when references to scientific articles are picked up. We analyze to which extent the reference structure in Wikipedia articles can be explained by two field perspectives, namely focusing on secondary accounts and on citation impact. This paper is part of a larger pilot study on Wikipedia as a lens into innovation-in-the-making (Simons et al. 2021).

Materials and Methodology

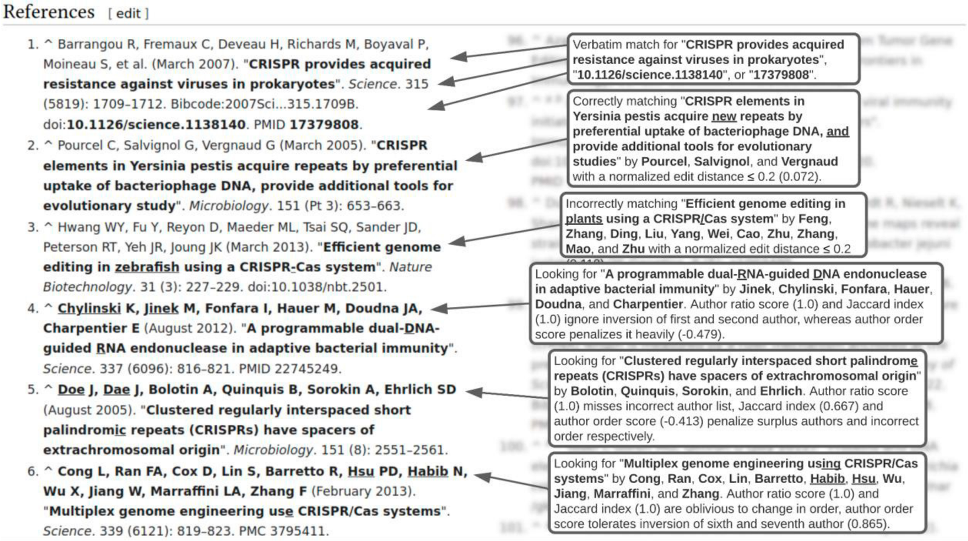
We provide two corpora of publications on CRISPR, the CRISPR Accounts Corpus and the CRISPR WoS Corpus, each representing a different perspective on the CRISPR innovation, and we compare them with the revision history of the central Wikipedia article on CRISPR.[i,ii] Both corpora are exploratory, and, to a certain extent, pragmatically defined, as patterns for literature referencing in Wikipedia in the specific case of innovations are as of yet unknown.

The CRISPR Accounts Corpus comprises publications referenced in predominantly secondary literature formats, such as reviews and short communications. These accounts present and discuss the development of CRISPR, thus conveying experts' and stakeholders' perspectives on which publications have been relevant for the innovation process—collectively called 'accounts'. To create this corpus, we searched for the phrases "crispr history", "crispr development", and "crispr discovery" on Google Scholar, and, based on Google's relevance ranking, reviewed the paginated search result pages until no more relevant publications were identified in a row of consecutive pages. 12 publications, despite containing history sections, were discarded for not focusing on the history of CRISPR, one for not being listed in WoS. The resulting 29 sources have been complemented by three web resources presenting CRISPR timelines, which were obtained by searching for "crispr timeline" on Google, and reviewing the search results pages accordingly. We extracted all references from the 29 sources through WoS and extracted references manually in case of the timeline documents.

The CRISPR WoS Corpus is based on a bibliometric field delineation of CRISPR in the Web of Science (WoS). Publications containing "crispr"—as a highly distinctive term—in the title are defined as the core field. As a second layer, we add publications containing "crispr" in their abstracts. For a third layer, representing influences and effects, we delineate publications for which the proportion of references or citations to the core field to total references or citations is higher than 30 percent, thus representing a substantial connection. The number of publications in this corpus amounts to 20,532.[iii] In order to represent an impact perspective in contrast to that of the history-oriented accounts, we sort this corpus by the absolute citation counts and cut off at 500 publications. This number is roughly in line with the magnitude of the references in the article (see below). The reason for not using citation windows here is rooted in the specific dynamics of an innovation, where early publications typically accumulate significant citation numbers over time[iv].

For both corpora we downloaded the metadata from WoS, resulting in **1,186** and **500** unique publications, respectively, to be matched against all revisions of the English CRISPR article. All revisions of the CRISPR article were downloaded using the MediaWiki API, collecting the HTML versions of each revision's page (2,072 revisions as of February 28, 2021). Each revision has a unique revision timestamp, text sections, such as headings, paragraphs, captions, tables, and lists, the reference sections *References* and *Further Reading*, as well as other metadata. While more recent revisions of articles usually apply Vancouver style to format references, Wikipedia does not have a single house style but expects the editors to adopt a consistent style within articles, potentially causing shifts of citation styles over the course of an article's history. This ambiguity together with the fact that editors easily introduce typos when manually adding a reference to an article necessitates the development of a fuzzy reference matching approach—at least we thought so at the outset.

Our reference matching approach implements heuristics with various degrees of precision, which can be divided into verbatim heuristics and fuzzy heuristics. As verbatim heuristics, we match titles, DOIs, and PMIDs of the publications against the entire article text including all references. All strings are converted to lowercase ASCII, with title matching additionally utilizing alpha-numerical normalization. As fuzzy matching heuristics, we match titles against extracted references and allow for a normalized edit distances of 0.2, 0.3, and 0.4, and combine the latter with three author matching strategies. We use the publication-to-reference-author ratio, the Jaccard Index of publication and reference authors, and an author order score: Each author of the publication is assigned a gain equal to its position in the inverted list of authors divided by its actual position, with the sum of all values being the ideal score. The authors of a given reference are then evaluated in turn, winning the same value if matching the author of the publication in the respective position and losing that value if not. The author order score is calculated by dividing the sum of these values by the ideal score.



References [edit]

1. ^ Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. (March 2007). "**CRISPR provides acquired resistance against viruses in prokaryotes**". *Science*. 315 (5819): 1709–1712. Bibcode:2007Sci...315.1709B. doi:**10.1126/science.1138140**. PMID **17379808**.
2. ^ Pourcel C, Salvignol G, Vergnaud G (March 2005). "**CRISPR elements in Yersinia pestis acquire repeats by preferential uptake of bacteriophage DNA, provide additional tools for evolutionary study**". *Microbiology*. 151 (Pt 3): 653–663.
3. ^ Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh JR, Joung JK (March 2013). "**Efficient genome editing in zebrafish using a CRISPR-Cas system**". *Nature Biotechnology*. 31 (3): 227–229. doi:10.1038/nbt.2501.
4. ^ **Chylinski K, Jinek M, Fonfara I, Hauer M, Doudna JA, Charpentier E** (August 2012). "**A programmable dual-DNA-guided RNA endonuclease in adaptive bacterial immunity**". *Science*. 337 (6096): 816–821. PMID 22745249.
5. ^ **Doe J, Dae J, Bolotin A, Quinquis B, Sorokin A, Ehrlich SD** (August 2005). "**Clustered regularly interspaced short palindromic repeats (CRISPRs) have spacers of extrachromosomal origin**". *Microbiology*. 151 (8): 2551–2561.
6. ^ **Cong L, Ran FA, Cox D, Lin S, Barretto R, Hsu PD, Habib N, Wu X, Jiang W, Marraffini LA, Zhang F** (February 2013). "**Multiplex genome engineering use CRISPR/Cas systems**". *Science*. 339 (6121): 819–823. PMC 3795411.

Verbatim match for "**CRISPR provides acquired resistance against viruses in prokaryotes**", "**10.1126/science.1138140**", or "**17379808**".

Correctly matching "**CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies**" by **Pourcel**, **Salvignol**, and **Vergnaud** with a normalized edit distance ≤ 0.2 (0.072).

Incorrectly matching "**Efficient genome editing in plants using a CRISPR/Cas system**" by **Feng**, **Zhang**, **Ding**, **Liu**, **Yang**, **Wei**, **Cao**, **Zhu**, **Zhang**, **Mao**, and **Zhu** with a normalized edit distance ≤ 0.2 (0.110).

Looking for "**A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity**" by **Jinek**, **Chylinski**, **Fonfara**, **Hauer**, **Doudna**, and **Charpentier**. Author ratio score (1.0) and Jaccard index (1.0) ignore inversion of first and second author, whereas author order score penalizes it heavily (-0.479).

Looking for "**Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin**" by **Bolotin**, **Quinquis**, **Sorokin**, and **Ehrlich**. Author ratio score (1.0) misses incorrect author list, Jaccard index (0.667) and author order score (-0.413) penalize surplus authors and incorrect order respectively.

Looking for "**Multiplex genome engineering using CRISPR/Cas systems**" by **Cong**, **Ran**, **Cox**, **Lin**, **Barretto**, **Habib**, **Hsu**, **Wu**, **Jiang**, **Marraffini**, and **Zhang**. Author ratio score (1.0) and Jaccard index (1.0) are oblivious to change in order, author order score tolerates inversion of sixth and seventh author (0.865).

Figure 1: Examples showing verbatim and fuzzy heuristics; divergent data underlined.

For the time being we refrain from calculating exact recall values, since it requires manually reviewing all 2,072 revisions for possible matches. However, for an estimate of what might have been missed, we manually checked and deduplicated all titles, DOIs and PMIDs that had been extracted from all references throughout the article's entire revision history, resulting in 324 DOIs, 302 PMIDs, and 465 titles. The latter boil down to around 370 titles, which result in 331 unique WoS items. When mapped to our corpora, only two publications were missed by the reverse procedure, thus being almost completely in line with our matching. Around 40 items not indexed in WoS are mainly articles in popular scientific or technological journals and blogs, as well as clinical trials and patents.

Results

Table 1 shows for both publication corpora the absolute numbers of matched publications, the relative numbers in relation to the sample corpora, and the precision of each method, calculated by manually checking the respective publications and matched references. The fuzzy matching heuristics D, E, and F generally identify more publications than the verbatim ones; at the cost of precision. The comparably smaller number of matches of the fuzzy heuristics G, H, and I may be a result of the fact that author lists in the Wikipedia article are somewhat less well-maintained than identifiers and titles.

A revision cites a publication if any of the matching heuristics correctly flags the publication in the revision. The number of uniquely matched publications resulting from both corpora is 201. Table 2 (left) shows the mean and median delays for each method relative to the timestamp of the earliest correctly matched revision for both corpora. In 184 out of 1,186 and 138 out of 500 cases, respectively, the earliest match can successfully be identified using the three verbatim heuristics alone. Only in one case a publication is identified earlier using a fuzzy heuristic.

Table 1: Evaluation of the reference matching heuristics applied to the revision history of the CRISPR article, dependent on the two corpora of relevant CRISPR-related publications.

Reference Matching Heuristic	**CRISPR Accounts Corpus**			**CRISPR WoS Corpus**		
	Absolute	Relative	Precision	Absolute	Relative	Precision
Verbatim matching heuristics						
A Title	173	14.59%	1.000	130	26.00%	1.000
B DOI	178	15.01%	1.000	135	27.00%	1.000
C PMID	177	14.92%	1.000	135	27.00%	1.000
Fuzzy matching heuristics						
D Title edit distance ≤ 0.2	182	15.35%	0.984	137	27.40%	0.985
E Title edit distance ≤ 0.3	186	15.68%	0.952	140	28.00%	0.957
F Title edit distance ≤ 0.4	203	17.12%	0.847	152	30.40%	0.849
G Title edit distance ≤ 0.4 + author ratio score = 1.0	166	14.00%	0.976	126	25.20%	0.984
H Title edit distance ≤ 0.4 + author jaccard index ≥ 0.8	152	12.82%	0.987	118	23.60%	1.000
I Title edit distance ≤ 0.4 + author order score ≥ 0.8	162	13.66%	0.988	124	24.80%	1.000

Table 2: Left: Delay in days for each heuristic in relation to the earliest correct match. Right: Number of times a publication (rows) is matched earlier by another heuristic (columns).

Reference Matching Heuristic	**Mean Relative Delay in Days**		**Median Relative Delay in Days**	
	Accounts	WoS	Accounts	WoS
A Title	69	70	0.0	0.0
B DOI	101	112	0.0	0.0
C PMID	40	51	0.0	0.0
D Title edit distance ≤ 0.2	67	67	0.0	0.0
E Title edit distance ≤ 0.3	68	67	0.0	0.0
F Title edit distance ≤ 0.4	67	67	0.0	0.0
G Title edit distance ≤ 0.4 + …	161	185	0.0	0.0
H Title edit distance ≤ 0.4 + …	160	173	36.0	32.5
I Title edit distance ≤ 0.4 + …	145	157	36.0	32.5

Comparison of Reference Matching Heuristics									
	A	B	C	D	E	F	G	H	I
A		10	47	1	1	2	1	0	0
B	16		62	18	18	18	9	8	9
C	19	26		20	20	21	12	4	5
D	2	9	49		0	1	0	0	0
E	2	9	49	0		1	0	0	0
F	2	8	47	0	0		0	0	0
G	25	28	62	25	25	24		2	4
H	52	56	81	53	53	50	34		2
I	52	57	85	52	52	50	34	0	

While there is very little discrepancy in the number of matches between the verbatim methods, the delays in Table 2 indicate that the methods sometimes match at different times, e.g. PMIDs earlier than titles and DOIs. For the majority of entries, however, this does not matter, as can be seen from the median. The matrix on the right side of Table 2 indicates for each matching heuristic (rows) how often another heuristic (columns) correctly identifies a publication in an earlier revision. The matrix depicts the results based only on the CRISPR Accounts Corpus and only takes correct matches into account. Since all matching is aimed at eliciting the earliest revision, a relaxed method might incorrectly select a revision to be the first one for a specific publication, which leads to a lower overall recall if cleaned for correctness. The title edit distance heuristic with the most relaxed threshold of 0.4 therefore seems more sensible than the more stringent methods using thresholds of 0.2 and 0.3, being superseded by the PMID methods in only 47 rather than 49 cases. This, however, is an expected consequence of the former's overall weaker precision of 172 correctly identified publications, as compared to 179 and 177 for the latter methods. Overall, the results of the combination of verbatim methods can hardly be improved by the relaxed methods, even with regard to delays.

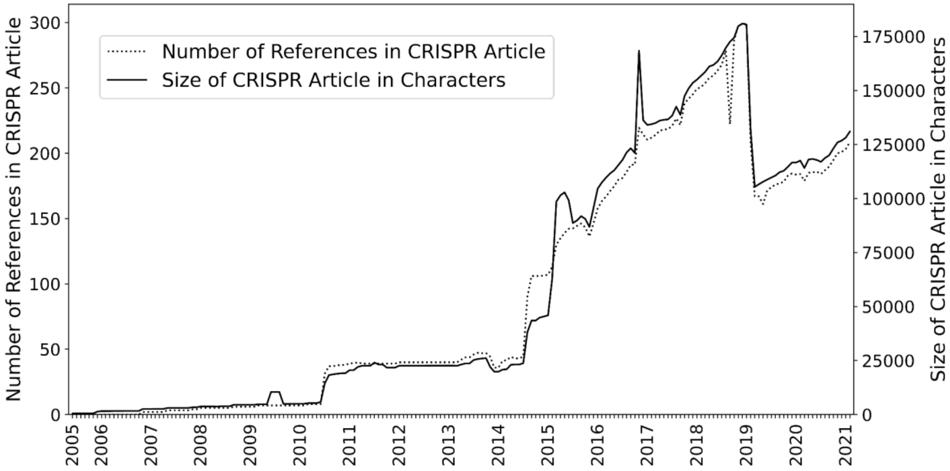


Figure 2: Smoothed dynamics of text growth of CRISPR article to growth of references.

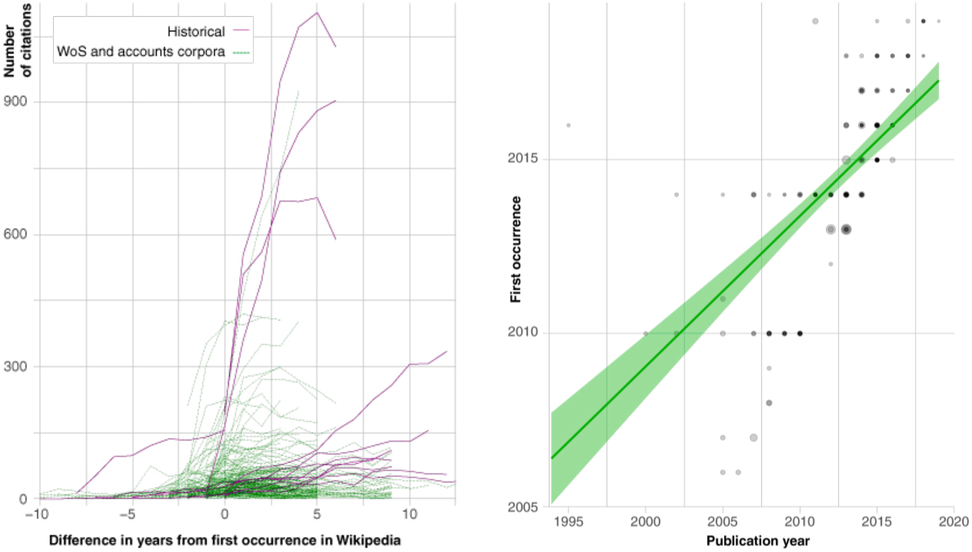


Figure 3: Citation distributions in relation to occurrence in Wikipedia. *Left:* Matched publications (201) in yearly citation counts, with date of first occurrence in CRISPR article as baseline (so that citations before date appear with negative numbers). Some prominent cases—on basis of

citations by secondary accounts and references' occurrence in current version of History section—marked by solid lines. *Right:* Publication years over years of first occurrence. Point sizes correspond to number of citations in absolute numbers, overlapping points displayed darker.

Both graphs in Figure 2 show a step in 2010 coinciding with the introduction of a 'History' section in the article. Key publications result from the years 2011 to 2013, which is reflected in the growth of text and references from 2014 onwards. In 2019, a section is moved to a new article on CRISPR gene editing. Similarly, Figure 3 on the right side shows that a number of publications dating from 2000 to 2010 were added in 2010 and a similar phenomenon can be observed in 2014, corresponding to the dynamics in Figure 2. These patterns suggest that the delays with which publications are referenced on the CRISPR page correspond to general dynamics of Wikipedia's edits on this topic. The graph on the right side of Figure 3 shows the first occurrence happens before the peak of the respective citation distribution in many cases.

Conclusion

We explored Wikipedia's usage of scientific literature in Wikipedia's article on CRISPR and the timeliness of its referencing patterns in order to gauge its relevance and adequacy as a medium for the representation and tracing of scientific innovations. More specifically, we proposed matching procedures to map from WoS publication corpora to all revisions of Wikipedia's central article on CRISPR. The results are promising: initial evidence suggests that substantial portions of the CRISPR/Cas9 literature referenced in Wikipedia are highly cited or have been acknowledged by experts in the field. We observe that the currency of referencing improves over time, and that article editing dynamics is captured as well. For the CRISPR/Cas9 case we can give evidence that a combination of verbatim matching heuristics yields sufficient accuracy, thus making Wikipedia an interesting object for analyses of science communication in addition to standard bibliometric sources.

References

Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P. T., Mavergames, C., & Gruen, R. L. (2014). Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. *PLoS Medicine*, *11*(2), e1001603.

Estevez, B., & Cukierman, H. (2012). The climate change controversy through 15 articles of Portuguese Wikipedia. Wikipedia Academy.

Garcia del Valle, E. P., Lagunes Garcia, G., Prieto Santamaria, L., Zanin, M., Menasalvas Ruiz, E., & Rodriguez Gonzalez, A. (2018). Evaluating Wikipedia as a Source of Information for Disease Understanding. 2018 *IEEE 31st International Symposium on Computer-Based Medical Systems* (CBMS), 399–404.

Giles, J. (2005). Internet encyclopaedias go head to head. Nature, 438(7070), 900–901.

Jemielniak, D., Masukume, G., & Wilamowski, M. (2019). The Most Influential Medical Journals According to Wikipedia: Quantitative Analysis. *Journal of Medical Internet Research,* 21(1).

Reavley, N. J., Mackinnon, A. J., Morgan, A. J., Alvarez-Jimenez, M., Hetrick, S. E., Killackey, E., Nelson, B., Purcell, R., Yap, M. B. H., & Jorm, A. F. (2012). Quality of information sources about mental disorders: A comparison of Wikipedia with centrally controlled web and printed sources. *Psychological Medicine*, 42(8), 1753–1762.

Simons, A., Kircheis, W., Schmidt, M., Potthast, M., Stein, B. (2021). Who are the Heroes of CRISPR? Priority Disputes on Wikipedia. Manuscript under review.

Zagovora, O., Ulloa, R., Weller, K., & Flöck, F. (2020). "I Updated the <ref>": The Evolution of References in the English Wikipedia and the Implications for Altmetrics. arXiv:2010.03083

[i] https://en.wikipedia.org/wiki/CRISPR

[ii] There are related Wikipedia articles (*Cas9* and *CRISPR Gene Editing*) but we limit our study to the central one.

[iii] On a frozen version of the WoS raw database from April 2020, to ensure reproducibility.

[iv] We also opted against field-normalization due to biases in case of high-impact multidisciplinary journals.

A Case Study of the Epistemic Function of Citations - Implications for Citation-based Science Mapping

Cara Seitz[1,] Marion Schmidt[2], Nathalie Schwichtenberg[2] and Theresa Velden[2]

[1] *cara.seitz@gmail.com*
IBI, Dorotheenstraße 16, 10117 Humboldt-University, Berlin (Germany)

[2] *schmidt@dzhw.eu, schwichtenberg@dzhw.eu, velden@dzhw.eu*
DZHW, Schützenstraße 6a, 10117 Berlin (Germany)

Abstract

The use of citations as indicators of topical relatedness of publications is common in the algorithmic mapping of the structure of science. References to source documents, however, may serve a variety of epistemic functions, and hence represent rather different dimensions of topical relatedness. They may pertain to the research methods used, the empirical objects studied, the theoretical resources build on, the research questions pursued, or the external motivation for and relevance of the work. In this case study, we explore the diversity in topical dimensions along which publications are linked in citation networks, by coding the epistemic function of in-text citations. To construct an informative sample of publications, we first made a quality assessment of an existing field delineation of invasion biology and identified the core and the boundary of the field. We then sampled 9 publications from each layer to analyze the epistemic functions of citations. Our preliminary results for the core layer show the diversity of epistemic functions that underlie in-text citations and suggest substantial variation between studies. While current approaches to science mapping use citation links as generic indicators of relatedness, future algorithmic extractions that distinguish between epistemic functions of citations might greatly enrich science mapping.

Introduction

The use of citation links as indicators of topical relatedness is common in the algorithmic mapping of the structure of science (see e.g. Sjögårde & Ahlgren 2018). The resulting structures, consisting of interrelated groupings of publications are commonly interpreted as scientific fields (in global mappings), and topics or subfields (in field mappings). However, very few validation studies are done that inspect such algorithmically generated structures in order to explore their interpretation and relate them to a sociological understanding of the social and cognitive structures of science (Gläser et al. 2017). There is a tacit understanding that topics structures overlap and publications may reference each other along different topical dimensions (e.g. with regard to the methods used versus the empirical objects studied). However, current mapping approaches use citations as generic indicators of topical relatedness and apply algorithms that produce disjoint, non-overlapping clusters of publications, without systematically validating how these pragmatic choices affect the representation of fields and topics by these maps.

To improve our understanding of how such methodological choices may influence the topic structures represented by algorithmically generated science maps, we take a closer look at citations as the input data of the algorithmic mapping process, and investigate the diversity of dimensions of topical relatedness that citations may incorporate. We proceed in two steps that both involve an intellectual assessment, first of publications, then of citation links: In a first step, we investigate how the publications in a lexically delineated field data set relate to the cognitive structure of the targeted field, distinguishing four ordinal classes of degree of 'field focus.' This is to deepen our understanding of how the publications that we select for citation analysis in the next step relate to the targeted research specialty. We then construct a purposive sample of publications to represent different types of studies in the targeted research specialty and investigate the variation in topical dimensions of relatedness that citations entail by zooming in on their epistemic function. Early work in citation classification has primarily examined the informative value of citations for assessing the impact of a cited publication (e.g. Moravcsik and P. Murugesan 1975, and Chubin & Moitra 1975). More recent work in citation classification has taken on further objectives, such as improving literature retrieval systems and automatic article summarization (Jha et al. 2017, Teufel et al. 2006). Our interest here is to consider the epistemic function of a citation in the citing text. Our coding scheme is inspired by recent work in the sociology of science that distinguishes different forms of epistemically mediated interdependencies in the collective knowledge production in the sciences (Gläser et al. 2018).

We present preliminary results, since the second half of the citation classification is ongoing. Our results about the degree of field focus for a random sample of 100 publications from the field data set are not only relevant for the construction of the purposive sample for this study, but also provide insights into the quality of the field delineation of invasion biology conducted by Held & Velden (*in preparation)*. We further report proportions of epistemic functions of in-text citations across different types of studies in invasion biology, and show how the proportion of cited sources that are located inside or outside the field differ by epistemic function.

Data and Methods

This is a case study, focused on publications in the field of invasion biology. Invasion biology is a research specialty that we are also studying through ethnographic observations and expert interviews to better understand research and sharing practices in this field. This background informs the sampling approach that we adopt in this study, and the decisions it entails, namely to distinguish publications by their degree of focus on the target field (invasion biology), and by study type, in order to stratify the sample by types of studies that are common in the field.

Field Data Set

To sample publications from the field of invasion biology we start from a bibliometric field data set that was generated using a lexical query to retrieve relevant articles and reviews from the Web of Science SCI and SSCI databases (Held & Velden, *in preparation*). The lexical query is an extension of the original query by Vaz et al. (2017) that was used in a previous citation-based mapping of the field of invasion biology (Held & Velden 2019). After constructing the direct citation network of the publications retrieved using the UTs of references, the final field data set is defined as the giant component of the network[i]. It consists of 53,524 publications published between 2000 and 2019.

Field Focus Coding

Since the manual coding of in-text citations is time consuming, we had to restrict ourselves to coding a small number of publications. To ensure the sample would represent studies in invasion biology well, we randomly sampled 100 publications from the field data set that were published in 2018, and classified them by the degree of their research focus on invasion biology ('field focus'). We defined a coding scheme that distinguishes four degrees of field focus, described in table 1. Coding decisions were taken in a step-wise process: first by determining whether a link to invasion biology existed or not[ii]. Then, if a link existed, by determining whether the link was strong or not. And finally, if the link was strong, by deciding whether the focus was singularly on invasion biology, or not. Following this decision tree, the 100 publications were coded independently by two coders. Basis for the assessment was the way that authors frame the contribution of their publication by composing title and abstract, choosing author keywords and selecting a journal to publish in. The inter-coder reliability, measured using a linearly weighted Cohen's Kappa, was 0.72[iii], a value which is commonly regarded as indicative of substantial agreement.

Table 1. Coding Scheme for Field Focus

Code	*Definition*
Core	Singular focus on invasion biology.
Boundary	Additional research focus, outside invasion biology.
Periphery	Explicit contribution to invasion biology is marginal.
Outside	No discernible connection to invasion biology.

Purposive Sample Construction for Citation Coding

We include in our sample core and boundary publications. Combined, we would argue, these types of publications define the field of invasion biology. We further take into account the type of study, distinguishing 'empirical' (observational, experimental), 'modeling'[iv], 'theoretical', and 'application-oriented' studies, in order to reflect the great variation in research approaches in invasion biology. We found the large majority of core and boundary publications to be empirical studies. Because theoretical studies were rare and occurred only among core publications, we omitted them from our sample for citation coding, and stratified the final sample of 9 core and 9 boundary publications by three study types: application-oriented, empirical, and modeling.

Citation Coding

Our classification scheme for capturing the epistemic function of in-text citations distinguishes between five epistemic functions that in-text citations may serve when referring to a source: clarifying the methodological approach (*method*), explicating properties of the empirical object (*empirical object*), drawing on or disputing theoretical insights (*theory*), defining and discussing the research problem (*research problem*), and referring to the external motivation or practical application of the work (*ext. interest*). If the context was ambiguous, we coded an in-text citation with up to two functions. We tried to use this option sparingly. The coding was conducted by two coders, one coding the core set of publications (see Seitz 2020), the other coding the boundary set, which is ongoing work. For resource reasons we could not afford double coding the entire sample. To ensure coding consistency, we performed selective double coding and review.

First Results

Field Membership

Our assessment of the field focus of randomly selected publications in the field data set finds that 71% (± 9%)[v] directly pertain to the field of invasion biology, either as core publications (49% ± 10%), or as boundary publications (22% ± 8%). Twenty-three percent (± 8%) are only peripherally related (e.g. simply mention the problem of invasive species as an ecological concern among other concerns). The remaining 6% (±5%) of publications are erroneously included, with no obvious relationship to the field.

Epistemic Functions of Citations

We observe considerable variation between the 9 core invasion biological studies in our sample. As figure 1 shows, the epistemic functions of in-text citations vary between studies, showing strong variation even among studies of the same study type. The most suggestive difference between study types seems to be that application-oriented studies tend to have more in-text citations that relate to external interest than empirical or modeling studies. Overall, in the aggregate, external interest and theory related in-text citations are fewer than empirical object, method or research problem related citations.

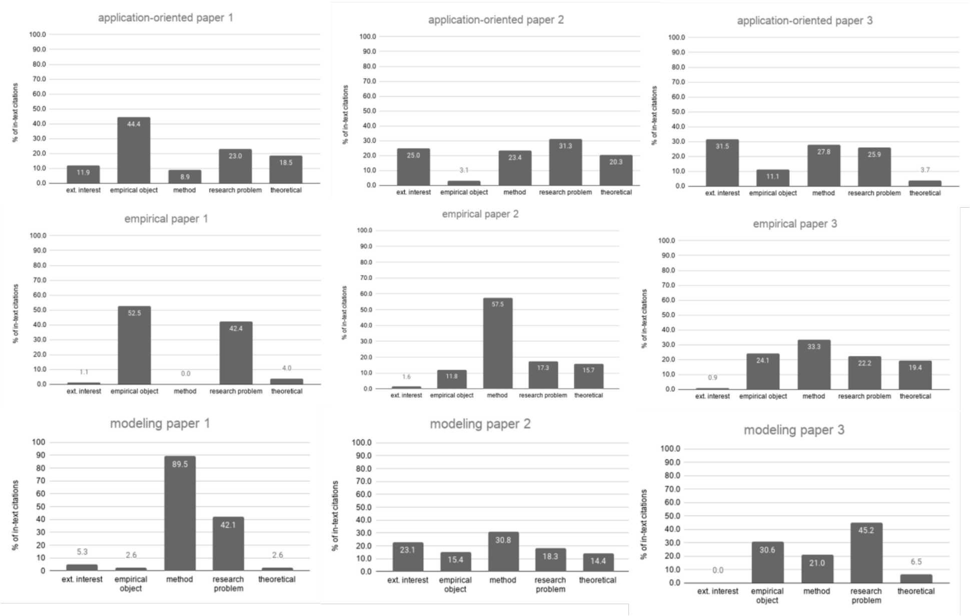


Figure 1. Epistemic function of in-text citations for each of the 9 publications in the core invasion biology sample.

Figure 2 shows the extent to which in-text citations in the 9 core invasion biology publications refer to publications inside versus outside the field data set. Only cited publications carrying a WoS UT and published within the time window of the field data set (2000-2019) are considered. Of the 568 references, 354 references in the 9 sample articles fulfill this criterion, corresponding to 571 in-text citations. Sources outside the field data set dominate among method related citations, whereas citations related to research problem, theory, or application aspects predominantly refer to publications within the field data set. For empirical object related citations the proportion of internal versus outside sources is almost even.

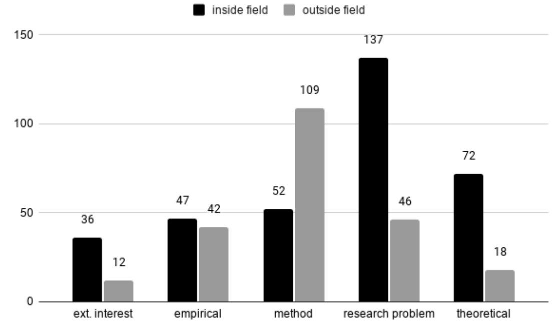


Figure 2. Proportion of in-text citations to publications inside versus outside the field data set by epistemic function (for subset of publications indexed by WoS and published in 2000-2019)

Discussion

How to get field delineation right is a critical issue in the bibliometric study of scientific fields (Zitt et al. 2019). Our coding of field focus is one possible approach to validate results and reveals that an estimated 71% of the lexically delineated field data set consists of publications that can be considered to represent research in invasion biology. Since this kind of validation exercise is rare, there is limited data to compare our result to. Haunschild et al. (2018) probed the purity of an algorithmically generated field classification by intellectual inspection of a random sample of 123 publications. They found roughly 50% of publications in the sample to belong to the target research specialty of 'common water splitting'. The decision criteria for whether a publication is considered to belong to the target field are not made explicit in their publication. Our approach differs in that we define a more fine-grained classification along with coding guidelines that help distinguish between those classes.

The process of developing the coding guidelines forced us to make explicit and agree on what, for us, defines field membership of a publication. By distinguishing between core and boundary publications, we found that a little more than 2/3rd of the field to consist of publications with a singular focus on invasion biology, and a little less than 1/3rd of the field to consist of publications that share a research focus on invasion biology with a research focus on a neighboring field. Importantly, some of the publications that were classified as singularly focused on invasion biology, simultaneously contribute to a field that can be interpreted as 'pervasively' overlapping. For example, invasion biological studies that examine the interaction of an invasive species with a potential biological control agent to be used in invasion control can be simultaneously regarded as studies in the field of pest control. This observation underlines the poly-hierarchical structure of scientific fields (Havemann et al. 2017).

The results of coding the epistemic functions of in-text citation in the 9 core invasion biology papers are intriguing. That application-oriented studies would refer more strongly to an external interest such as the motivation for the study and the application relevance of its result, seems plausible. The strong within group variation in the epistemic function of in-text citations is striking. It points to shortcomings of the high-level study classification that we define and points to a diversity of subtypes. Take e.g. the variation in the proportion of method related in-text citations among empirical studies. Empirical paper 1 constitutes a classical field study reporting a small-scale observational survey of species occurrences. It has no method related in-text citations. By contrast, empirical paper 2 uses genetic analysis to determine invasion success factors and proposes a methodological innovation, criticizing previous analyses for a methodological deficiency. In this paper more than 50% of in-text citations relate to methods.

If we look at the epistemic function of citations to the concurrent literature inside versus outside the field, this case study further provides a window into the epistemic interdependencies between fields. Our (small sample of) data suggests that invasion biology is highly reliant on method-related knowledge published in other fields, followed by knowledge about empirical objects. Knowledge pertaining to the research problem, external interest, and theoretical assumptions, on the other hand, is primarily referenced by citing publications from inside the field. Whether this pattern holds for other fields is an open question.

Conclusions

We provide empirical evidence for the diversity in topical dimensions of relatedness that are represented by in-text citations. We find great variation in the epistemic functions that drive citations to other publications within the field versus outside the field. The current practice in science mapping of using citations generically as signals of topical relatedness conflates these different dimensions that could be highly informative for science mapping, if only these different dimensions were algorithmically accessible for automated extraction.

References

Chubin, D. E., & Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? Social studies of science, 5(4), 423-441.

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. Scientometrics, 111(2), 981-998.

Gläser, J., Laudel, G., Grieser, C., & Meyer, U. (2018). Scientific fields as epistemic regimes: new opportunities for comparative science studies. Technical University Technology Studies Working Papers TUTS-WP-3-2018

Haunschild, R., Schier, H., Marx, W., & Bornmann, L. (2018). Algorithmically generated subject categories based on citation relations: An empirical micro study using papers on overall water splitting. *Journal of Informetrics*, *12*(2), 436-447.

Havemann, F., Gläser, J., & Heinz, M. (2017). Memetic search for overlapping topics based on a local evaluation of link communities. Scientometrics, 111(2), 1089-1118.

Held, M., & Velden, T. (2019). How to interpret algorithmically constructed topical structures of research specialties? A case study comparing an internal and an external mapping of the topical structure of invasion biology. ISSI, 1933–1939.

Jha, R., Abu-Jbara, A., Qazvinian, V., & Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. Nat. Lang. Eng., 23(1), 93-130.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. Social studies of science, 5(1), 86-92.

Seitz, C. (2020) Epistemic Functions of Citation: A Comparative Analysis of Different Functions of Citations and Their Implications for Modelling Citation Networks. Masterthesis, Humboldt University, Berlin, Germany.

Sjögårde, P., & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics*, *12*(1), 133-152.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006, July). Automatic classification of citation function. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 103-110).

Vaz, A. S., Kueffer, C., Kull, C. A., Richardson, D. M., Schindler, S., Muñoz-Pajares, A. J., ... & Honrado, J. P. (2017). The progress of interdisciplinarity in invasion science. *Ambio*, 46(4), 428-442.

Zitt, M., Lelu, A., Cadot, M., & Cabanac, G. (2019). Bibliometric delineation of scientific fields. In Springer Handbook of Science and Technology Indicators (pp. 25-68). Springer, Cham.

[i] Held &Velden (2019, in preparation) disregard publications outside the giant component in order to increase the precision of the field delineation.

[ii] At times, publications picked up by the lexical query do not relate to invasion biology. This happened when a publication uses a term in the abstract that is part of the lexical query in order to signal invasiveness (such as 'introduced'). To be relevant, the lexical query requires such terms to be combined with another relevant term (such as 'species'), however it does not restrict the order of terms and permits a two word distance between terms, which can lead to irrelevant matches.

[iii] After the inter-coder assessment, one item in the coding guidelines was tweaked one last time in order to increase the internal consistency of the guidelines. The change meant to categorize studies that have an invasive species as empirical object, but do not articulate a link to an invasion biological research question, as peripheral – in agreement with the weight given in our qualitative classification approach to the framing of studies. This resulted in the reassignment of three studies that focus on invasive species from boundary to peripheral. We did not re-assess inter-coder reliability after this change.

[iv] Studies that use computational or statistical modelling approaches to predict the distribution of invasive species or factors that influence mechanisms of invasive processes.

[v] Based on a 95% confidence interval for the estimate of a population proportion.

Geographic Differences in the Uptake of Open Access

Marc-André Simard[1], Gita Ghiasi[2], Philippe Mongeon[3] and Vincent Larivière[4]

[1] *marc-andre.simard.1@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, 3150 rue Jean-Brillant, Montréal, Qc (Canada)

[2] *gita.ghiasi.hafezi@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, 3150 rue Jean-Brillant, Montréal, Qc (Canada)

[3]*pmongeon@dal.ca*
School of Information Management, Dalhousie University, Rowe Management Building, Suite 4010, 6100 University Avenue, Halifax, NS (Canada)

[4]*vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, 3150 rue Jean-Brillant, Montréal, Qc (Canada)

Abstract

Open Access (OA) is a process that aims to make research output freely available on the internet. The OA movement originated from the growing demand to make research more accessible worldwide. It has been gaining a lot of momentum, with the implementation of several OA policies by funding institutions and the development of several new platforms that facilitate the publication of OA content at low cost. Studies have shown that nearly half of the scientific literature could be available online for free, but only a few have compared the use of OA literature at the country level and from a worldwide perspective. Along these lines, this study aims to provide a global picture of the current state of OA adoption by countries, using two indicators: publications in OA and references to articles in OA. We find that, on average, low-income countries are publishing and citing OA at the highest rate, while upper middle-income countries and higher-income countries publish and cite OA articles at below world-average rates. These results highlight national differences in OA uptake and suggest that more OA initiatives at the institutional, national and international levels are needed to support a wider adoption of open scholarship.

Introduction

Open Access (OA) is a process that aims to make research output freely available on the public internet, allowing the users to read, download, copy, distribute, print, search or link to the full text without any financial, legal or technical barrier, in accordance with an open copyright license. The OA movement originated from the increasingly growing demand to make research more available worldwide. One of its milestones was 2001's Budapest Open Access Initiative, which established the first clear distinction between the two main types of OA: self-archiving (green OA) and a brand-new generation of journals that would allow scholarly material to be distributed in its full form, freely on the publisher's website (gold OA). Gold open access journals may be associated with charge article processing charges, originally intended to cover publication costs. The movement has recently been gaining momentum, with the implementation of several OA policies by funding institutions worldwide and the development of several OA new platforms such as including Open Journal System (OJS), SciELO, and Érudit. These platforms facilitate the publication of OA content at low cost. According to different studies, nearly half of research articles could be available online at no cost (Archambault et al., 2014; Piwowar et al., 2018; European Commission, 2019). Previous studies have addressed different aspects of the OA phenomenon, such as the availability of articles (Archambault et al., 2014; European Commission, 2019; Piwowar et al. 2018), the so-called "citation advantage" of OA articles (Antelman, 2004; Swan, 2010; Piwowar et al., 2018), and OA mandates (Larivière & Sugimoto, 2018). However, very few articles (Evans & Reimer, 2009; Iyandemye & Thomas, 2019) have compared the use of OA literature at the country level. In this paper, we provide an up-to-date portrait of the adoption of OA across the world, distinguishing between publishing in OA and citing OA publications, between two types of OA (green and gold).

Research Objectives

The purpose of this paper is to provide a global picture of the current state of OA adoption by countries, using two indicators: publications in OA and references to articles in OA. Our research questions are as follows:

1. What proportion of the scientific output of countries is Open Access?
2. To what extent are different countries citing literature in open access?
3. Does OA publications correlate with OA references at the country level?
4. How does OA usage relate to the level of income of countries?

Methods

We collected all articles and reviews in the Web of Science (WoS) published between 2015 and 2019. Science being increasingly collaborative, publications are generally authored by multiple individuals, often from different institutions and countries, among whom the work has typically been unevenly distributed (Larivière et al. 2015). We argue that choices in terms of cited literature and publication venues are predominantly made by authors who played a leading role in the research. Thus, when assigning a publication to a country, we only use the institutional affiliations of the first and corresponding authors. We further classified countries based on the World Bank's country classifications by income level (2018–2019).

To determine the OA availability and OA type of a paper or reference, we searched for its DOI in the Unpaywall (http://unpaywall.org) database, which identifies OA content from open indexes (e.g., Crossref, DOAJ), journals and repositories. This data indicates whether the paper is published in an OA journal (gold OA) and if a publication is found in a repository (Green OA). These two categories are not mutually exclusive since a paper can be deposited in a repository even if it is published in an OA journal. Because we use the DOI to link the WoS and Unpaywall data, WoS publications that do not include a DOI are excluded from our study. The distribution of publications by fields may differ between countries. We weight this field-normalized indicator based on the proportion of publications in each field by that country. The resulting indicator is a weighted and field-normalized measure of OA usage, which we apply to the analysis of OA publications and references. The main purpose of our analyzes being to identify differences in OA uptake between countries, we add another layer of normalization comparing the weighted and field-normalized indicator to the world average (which is represented by a value of 1 in the results).

Results

Tables 1 and 2 presents an overview of our dataset by discipline and type of OA in terms of publications and references, respectively. In table 1, we see that the proportion of publications in OA varies by fields, with 40.5% of papers being OA in biomedical research (BM; 40.5%) and 20.2% and 20% in natural sciences and engineering (NSE) and in social sciences and the humanities (SSH) respectively. The very small proportion of papers that are in gold OA only (1.3 - 3.9%) shows the large overlap between green and gold OA.

In table 2, we see that the share of references to OA papers also varies by fields with Biomedical Research citing more OA papers. However, we can observe that the proportion of references to gold OA articles is generally lower than the proportion of gold OA publications, with social sciences and humanities being the only exception. But overall, the proportion of OA references is higher than the proportion of OA publications. Again, the small proportion references to articles that are available only via gold OA shows the large overlap between green and gold.

Table 1. Proportion of OA publications by type and by fields.

Field	*Number of publications*	*% OA*	*% Gold OA*	*% Green OA*	*% Gold Only*	*% Green Only*
BM	2,529,405	40.5%	18.9%	39.3%	1.3%	21.7%
NSE	3,371,682	20.2%	7.2%	16.2%	3.9%	13.0%
SSH	757,466	20.0%	6.8%	18.5%	1.5%	14.1%
All	6,658,553	27.9%	11.4%	25.2%	2.7%	16.4%

Table 2. Proportion references to OA by type and by fields.

Field	*Number of references*	*% OA*	*% Gold OA*	*% Green OA*	*% Gold Only*	*% Green Only*
BM	14,913,746	48.2%	13.8%	47.5%	0.6%	34.4%
NSE	18,471,269	24.9%	6.0%	23.5%	1.5%	19.0%
SSH	2,055,463	32.2%	7.7%	31.6%	0.6%	24.4%
All	35,440,478	35.1%	9.3%	34.1%	1.1%	25.8%

Figures 1 and 2 reveal that countries mostly cite OA more often than they publish in OA. Sub-Saharan African countries publish and use OA more often. In North America, the United States publishes and cites OA more than the world average, however, Canada publishes in OA less often but cite OA more often. South American countries also cite OA more often, however, Brazil publishes more in OA rather than citing it. In Europe, Western European countries mostly publish and cite OA, while the trend is opposite for the Eastern European countries. In Asia, North Africa and the Middle East, most countries publish or cite OA less often.

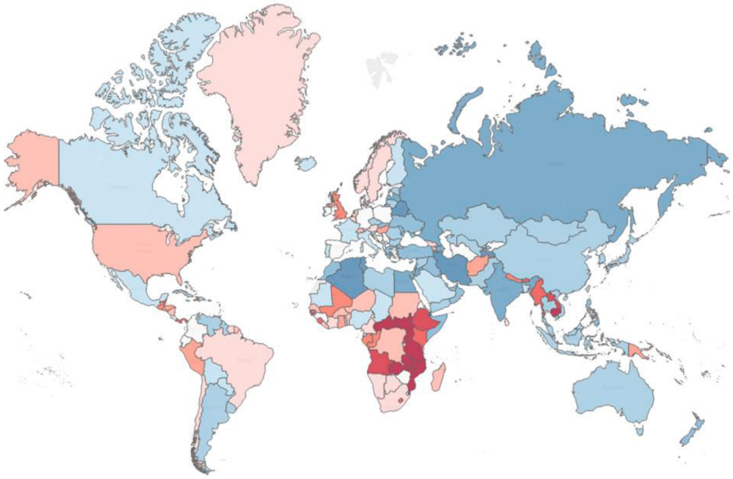


Figure 1. Weighted and normalized map of number of OA publications by country. Red indicates that a country is above the world average, blue indicates it is below the world average. White represents the world average.

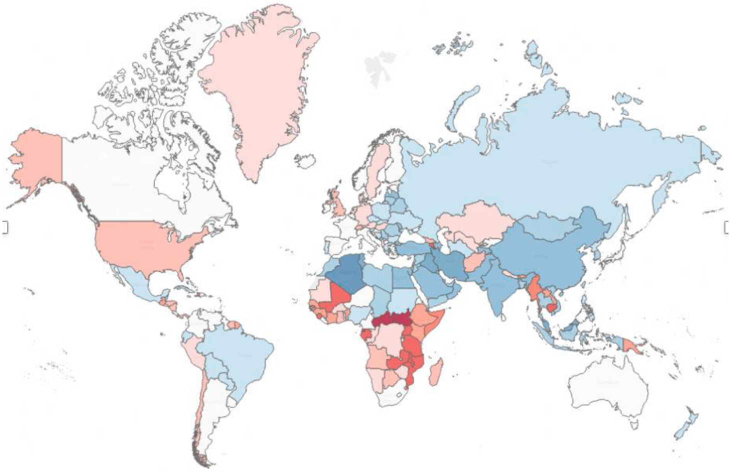


Figure 2. Weighted and normalized map of references made to OA publications by country. Red indicates that a country above the world average, blue indicates it is below the world average. White represents the world average.

Figure 3 illustrates the relationship between the weighted and normalized OA publication indicator of countries with the weighted and normalized reference indicator. A Pearson correlation shows a very strong relationship between the two indicators (r = 0,870; p < 0,001), which indicates that the more a country publishes in OA, the more it is likely to make references to OA papers.

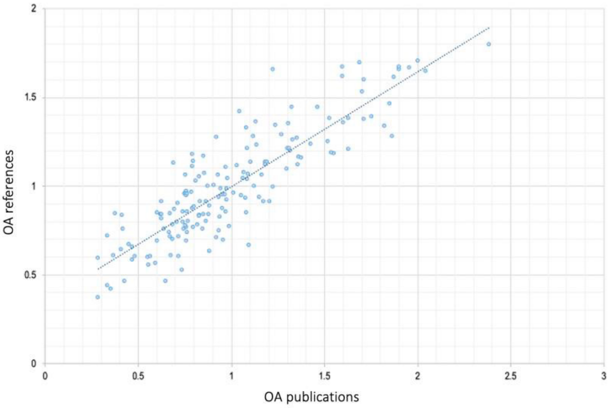


Figure 3. Scatter graph indication the relation between the weighted and normalized OA publication indicator of countries with the weighted and normalized reference score.

Figure 4 shows plots of countries according to their OA publications and references compared to the field-normalized world average. We observe that overall, as figure 1 suggested, the two indicators correlate. The correlation is, however, 1.3 times stronger for lower middle-income and low-income countries and upper-middle-income countries. On average, the different groups are positioned in different quadrants: the higher and upper-middle-income countries are on average in the third quadrants while low income and lower middle-income countries are situated in the first quadrant. This shows that overall, countries with higher income have a lower tendency to both publish and cite OA papers than lower income countries.

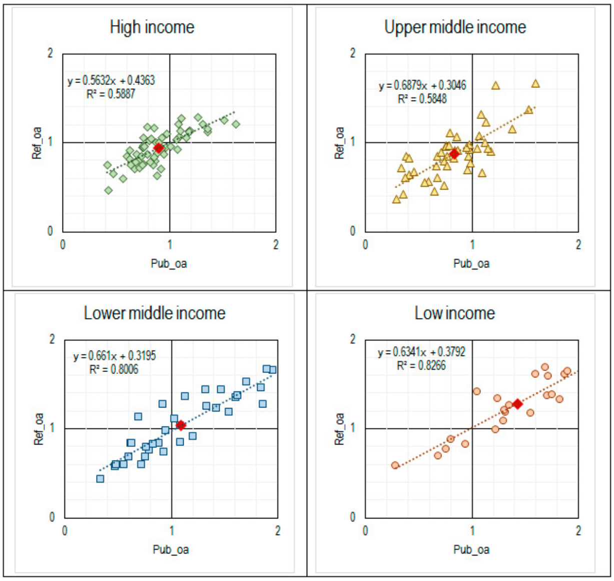


Figure 4. OA publication vs. references per country income (red dots; average for all the countries of the same income category).

Conclusion

Our findings show that Sub-Saharan Africa is publishing and citing OA at a higher rate than the rest of the world, while we found that the Middle East and Asia are the areas where the proportion of publications available in OA is lower, as their use of OA. This could be explained by the fact that APCs are mostly waived for Sub-Saharan countries (e.g., research4life countries) but is not or only partially waived for most Middle Eastern and Asian countries. One other possible explanation is that the national and transnational OA initiatives such as plan S for the European Union and PubMed Central in the United States are almost non-existent in these countries, and institutional repositories are not well developed. Looking at income levels, our findings also reveal that lower middle-income and low-income countries are those who publish and cite the most OA. Again, this may be partly explained by the fact that APCs are generally waived for these countries, making publishing in gold OA more accessible for them. Moreover, some commercially owned gold OA journals may have less strict publication criteria and given the lack of research infrastructure in these countries, they may be more likely to publish in these journals (Butler, 2013). We also found that the upper-middle-income countries behave similarly to the higher-income countries. However, the underlying mechanisms behind the use of OA potential may be different. For instance, upper middle-income countries may lack the resources to pay for APCs on top of high subscription prices for closed journals (both of which they don't qualify for waivers), and which is not necessarily the case for high-income countries. There may also be other factors, such as a lower reputation of OA journals in certain parts of the world. Ultimately, while high-income countries have mandates and repositories, and low-income countries have waivers, our results highlight national differences in OA uptake and suggest that more OA initiatives at the institutional, national and international levels are needed to support a wider adoption of open scholarship.

References

Antelman, K. (2004). Do Open-Access Articles Have a Greater Research Impact? *College & Research Libraries*, *65*(5), 372- 382.

Archambault, E., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebout, L. & Roberge, G. (2014). *Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Level—1996–2013* (RTD-B6-PP-2011-2). European Commission; Science-Metrix.

Butler, D. (2013). Investigating journals: The dark side of publishing. *Nature News*, 495(7442), 433.

European Commission. (2019). *Trends for open access to publications*. European Commission.

Evans, J. A. & Reimer, J. (2009). Open Access and Global Participation in Science. *Science*, 323(5917), 1025 - 1025.

Iyandemye, J., & Thomas, M. P. (2019). Low income countries have the highest percentages of open access publication: A systematic computational analysis of the biomedical literature. *PLOS ONE*, 14(7), e0220229.

Larivière, V., Gingras, Y., Sugimoto, C. R. & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, 66(7), 1323–1332.

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J. & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, 6, e4375.

Swan, A. (2010). The Open Access citation advantage: Studies and results to date. *University of Southampton Institutional Repository*. https://eprints.soton.ac.uk/268516/

From Citing Sentences to Causal Networks: The Causality Index

Henry Small

hsmall@mapofscience.com
SciTech Strategies, Inc., Bala-Cynwyd, Pa. (USA)

Abstract

Citing sentences for highly cited papers are examined to ascertain the presence of causal assertions. The incidence of causal statements varies by field of science, and the type of causality depends on the nature of the cited paper. Analysis focuses on papers in the biomedical category. Samples of citing sentences are analyzed to extract causes and their related effects. Aggregation of cause-effect patterns, following the unification of terminology, allows the construction of causal networks. These networks can be compared with the content of the cited paper to highlight new findings and points of consensus. Machine learning sheds light on the words associated with causal and non-causal assertions, and a grammatical analysis of causal sentences provides a way to partially automate the identification of statements of cause and effect by combining noun or noun phrase positions in sentences with the presence of causal words, leading to the formulation of a "causality index".

Introduction

According to Pearl (2018) we are all causal thinkers. It is natural for humans to try to make sense of our experiences by positing causes whether these explanations are scientific or conspiratorial. Philosophers from Aristotle to David Hume have developed philosophies of nature based on various types of causes (Bunge, 1959). While notions of causality may break down at the subatomic level, they remain a pervasive way of thinking about macroscopic phenomena in all branches of science from physical to social.

This paper came out of an investigation of the role of theory in citing sentences. Citing sentences were examined for samples of highly cited papers, and each paper was classified as dealing in some way with theoretical issues. On closer examination it was seen that discussions of theory were often embedded in statements of a causal nature where theories were cast as causes and their effects as empirical outcomes. This led to an attempt to extract causes and effects from the citing sentences and representing them as networks. As an example of how theories and causes are intertwined, consider how Newton's theory of gravitation "causes" the planets to travel in elliptical orbits around the sun. Here the theoretical construct of "gravity" is at least in part the cause of the observed phenomenon.

It turns out "causal networks", or "causal maps", have had a long history particularly in fields such as management, operations research, and political science (Narayanan & Armstrong, 2005; Axelrod, 1976). These efforts have also been described as "cognitive maps", "fuzzy causal maps" and "Bayesian causal networks." Data sources are commonly questionnaires, surveys or texts, but rarely scientific or technical texts (Sobrino, Olivas & Puente, 2010). Methods for extracting causal content are mostly qualitative and subjective, although increasingly algorithms utilizing deep learning are being developed (Li, et al. 2021; Trieu et al. 2020).

The advantage of performing this analysis on citing sentences is that we can tap into the multiple perspectives of citing authors on the unique causal aspects of a single cited paper, not only the viewpoint of the cited author. Linking a cited paper to its citing sentences also reveals the degree of agreement among citing authors on what causes what, and these diverse views can be aggregated to form a directed network of assertions where each arrow points from a cause to an effect. The resulting network can be interpreted as a collective model of the theory including, in some cases, its empirical outcomes.

Of course, not all cited papers and not all citing sentences are causal in nature. Sentences can also be descriptive, procedural or programmatic. Method papers, for example, often contain procedural sentences as do sentences citing such papers. While methods appear at first glance not to be causal, their real causal basis may be hidden in earlier generations of papers. Interestingly, this situation is analogous to technologies that have predictable outcomes and a deterministic basis which is hidden as in a "black box."

Various types of causation can be seen in these data ranging from hypothetical, to deductive in mathematics, to deterministic in chemistry, to statistical in medicine or the social sciences. There is also a range of certainties associated with these mechanisms, and their degree of empirical confirmation can, at least in principle, be modelled by Bayes's theorem using conditional probabilities (Small, 2019; Zeevat & Schmitz, 2015).

Data and Methods

The data for this study are drawn from the full text of papers published by Elsevier as described by Boyack, et al. (2018). The cited papers comprise the 2,500 most cited papers published between 2000 and 2004 which are separated into five broad fields of science. Citing sentences are taken from the same Elsevier full texts covering the years 2000 to 2016. The 500 papers for each field are divided into centiles.

The number of highly cited papers dealing with theoretical issues can be used as an approximate indicator of the fraction of causal citing sentences in each of the five fields. Fields with the highest fractions of causal citing sentences are biomedical sciences and social/behavioral sciences, followed by mathematics/computer science. Life science/geosciences and physics/engineering had the lowest fractions.

Table 1 shows examples of causal sentences drawn from various fields. In three of the five examples, the cause is identified with the subject of the sentence and the effect with the predicate. In the examples from biology and physical science this role is reversed due to the use of the passive voice signaled by verb - preposition combinations "caused by" and "due to".

Table 1. Examples of causal citing sentences from different fields.

Health
Smoking is the leading preventable **cause** of death in the United States.
Biology
Harmful heavy metal action in plants is **due to** the generation of reactive oxygen species and induction of oxidative stress.
Physical Science
The improved heat transfer by nanofluids is **caused by** the increased thermal dispersion due to the chaotic movement of nanoparticles.
Social Science
It has been shown that oil price changes **affect** economic growth asymmetrically.
Statistics
Maximum likelihood approaches have been shown to substantially reduce bias **arising** from missing data.

For the present study, causal analysis is performed only on the fifth centile of papers in the biomedical field. Focusing on the 5th rather than the 1st centile avoids the overrepresentation of method papers. The second most cited paper in the 5th biomedical centile (Caterina, et al., 2000), which is the test case below, has 764 citing sentences and deals with nociception.

Highly cited papers were classified as causal or non-causal based on the examination of a sample of about 20 citing sentences. On this basis roughly one-half of the papers in the 5th biomedical centile were deemed predominantly causal in nature. For these 50 papers, the samples of citing sentences were manually scanned to extract text strings representing causes and effects. The text strings were then joined together in directed networks as in Figure 1 representing each cited paper, where causes point to effects. The same node can be both a cause and an effect, as in the case of chaining, and causes can have multiple effects, and effects multiple causes.

Because citing authors often use different technical terms to refer to the same concepts or entities, it is necessary to unify the terminology in constructing these networks. The causal network for the Caterina paper is shown in Figure 1. The node labelled "TRPV1 knockout mice" is unified with "TRPV1 -/- mice", "TRPV1-deficient mice", and other variants. Each link in the network is labelled by the relative frequency of citing sentences linking the cause and effect and the links are labelled as having a "promoting" (+) or "inhibiting" (-) influence as is customary in the causal network literature. The network is also annotated to indicate effects having empirical evidence (italics), "new" relationships absent in the original cited paper (underlined text), and future possibilities (dashed outlines).

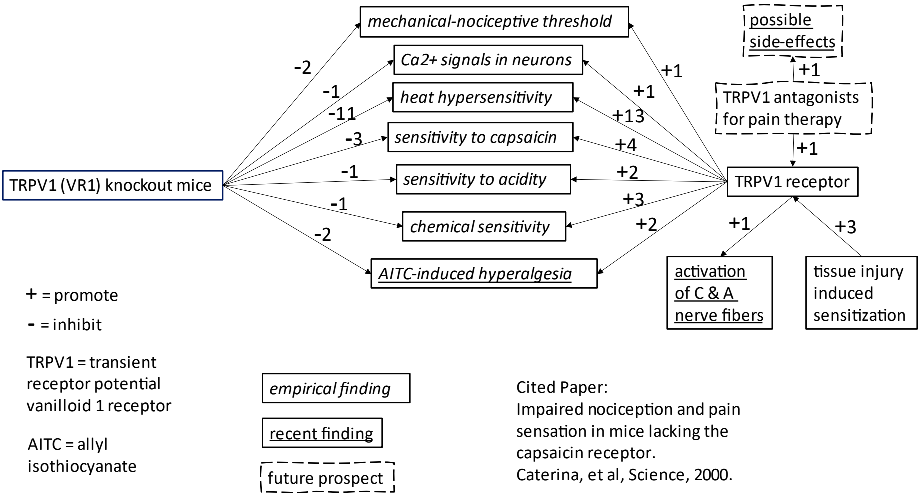

TRPV1 (VR1) knockout mice
mechanical-nociceptive threshold
Ca2+ signals in neurons
heat hypersensitivity
sensitivity to capsaicin
sensitivity to acidity
chemical sensitivity
AITC-induced hyperalgesia
-2
-1
-11
-3
-1
-1
-2
+1
+1
+13
+4
+2
+3
+2
possible side-effects
+1
TRPV1 antagonists for pain therapy
+1
TRPV1 receptor
+1
+3
activation of C & A nerve fibers
tissue injury induced sensitization
+ = promote
- = inhibit
TRPV1 = transient receptor potential vanilloid 1 receptor
AITC = allyl isothiocyanate
empirical finding
recent finding
future prospect
Cited Paper:
Impaired nociception and pain sensation in mice lacking the capsaicin receptor.
Caterina, et al, Science, 2000.

Figure 1. Causal network of citing sentences for the Caterina, et al. (2000) paper.

The following citing sentence illustrates the procedure used to construct the network: "The most salient phenotype of TRPV1 null mice is a profound loss of the thermal hypersensitivity normally associated with inflammation." The cause in this sentence corresponds to the node labelled "TRPV1 knockout mice" at the left of the Figure, while the effect corresponds to the node labelled "heat hypersensitivity." Clearly the challenge is to automate the construction of the causal network as far as possible.

To this end a machine learning exercise was undertaken using the Caterina paper as a test case. A sample of 327 sentences were selected from papers citing the Caterina paper, and manually classified as causal or non-causal. The Scikit-learn package was used for machine learning (Pedregosa et al., 2011). The median of the ten classifiers was 73% (F1 = .74). It is instructive to look at the coefficients of the individual words that define the optimal surface for a given classifier. For example, high coefficient words for the Bernoulli classifier included cause or effect words like "induced", "activation", "stimuli" and "responses" while low coefficient words were more action oriented like "performed" or "examined" but in general were more diverse. This suggests that, despite the modest accuracy, individual words can, to some extent, differentiate causal from non-causal sentences. Thus, a vocabulary of 77 words signalling causes or effects was compiled and used to code the sentences.

However, further evidence is required to make the distinction of causal from non-causal. This is provided by information on the relative positions of nouns and noun phrases within the sentence. It was observed that in some cases frequently occurring noun phrase pairs, such as "TRPV1 knockout mice" and "heat hyperalgesia", correspond to links on the map. Further, the nouns or noun phrases representing causes were most often associated with the subject of the sentence while effects were associated with the predicate (Findler & Bickmore, 1996). For example, based on the sample of sentences used to create Figure 1, "causes" appeared in the subject of the sentence 73% of the time and "effects" appeared in the predicate 87% of the time. This suggested the simple rule of parsing each sentence into segments defined by the position of the verbs. A rough division of subjects and predicates is thus obtained by applying a part-of-speech analysis to the citing sentence. An example is shown in Table 2 based on the simple citing sentence quoted above. Here the verb "is" separates the causal noun "TRPV1" in the subject of the sentence from the effect phrase "thermal hypersensitivity" in the predicate. For more complex sentences with multiple verbs, additional segments are defined and numbered sequentially. Part-of-speech analysis was implemented using the standard NLTK tagger with customizations for biomedical terminology (Bird et al. 2009), followed by separation of the sentence into numbered segments.

Table 2. Parts of speech (POS) for a citing sentence indicating cause (C) and effect (E).

Seq	*Word*	*POS*	*Cause/Effect*
1	The	DT	C
2	most	RBS	C
3	salient	JJ	C
4	phenotype	NN	C
5	of	IN	C
6	TRPV1	NNP	C
7	null	NN	C
8	mice	NN	C
9	is	VZB	E
10	a	DT	E
11	profound	JJ	E
12	loss	NN	E
13	of	IN	E
14	thermal	JJ	E
15	hypersensitivity	NN	E

Results and Conclusions

Combining information on the relative positions of noun phrases with one or more of the 77 cause/effect signal words noted above, it was possible to create a causality index for a given pair of nouns or noun phrases purported to be a cause-effect pair. First, we count the number of times the candidate noun-cause appears in a sentence segment that precedes the candidate effect-noun and divide this by the total number of times the two nouns appear jointly in sentences. Then we compute the percentage of sentences that contain signal words. For example, referring to Figure 1, a stem search is carried out across all citing sentences for the link representing "*TRPV1*" and "*heat*" where the first term appears in a segment prior to the appearance of the second term. This gives 157 sentences where the cause precedes the effect out of a total of 238 sentences where both terms are present, or 66%. Signal words appear in 142 of the 157 sentences (90%) and in 208 of the 238 sentences (87%). Since we want the causal word combinations to have a high causal signal as well as a high positional precedence, a suitable causality metric is the product of 66% and 90%, that is, the product of the proportion of sentences where the cause precedes the effect and the proportion of those same sentences containing causal signal words. A causality index of 100% would mean that the cause word precedes the effect word in all sentences, and all the sentences contain causal signal words.

Table 3 shows the analysis of a number of the links on Figure 1 where TRPV1 is the primary causal noun and the effect words correspond to specific links. Other non-causal effect words such as "studies" and "evidence" are included for comparison. The top six effect words with the highest causality indexes are represented on the map. Two effect words "antagonist" and "neuron" are on the map but have lower scores. The three words with the lowest scores are not on the map as effects. While the index looks promising as a means of distinguishing causal from non-causal word combinations, a comprehensive assessment would involve computing the index for all possible noun or noun phrase combinations in the citing sentences.

Table 3. Causality indexes for the cause word "TRPV1" combined with selected effect words.

Causality Index %	*Cause word*	*Effect word*	*On map*
63	TRPV1	hyperalgesia	Yes
60	TRPV1	heat	Yes
59	TRPV1	acid/proton	Yes
57	TRPV1	chemical	Yes
50	TRPV1	mechanical	Yes
38	TRPV1	capsaicin	Yes
29	TRPV1	evidence	No
24	TRPV1	antagonist	Yes
19	TRPV1	neuron	Yes
11	TRPV1	mice	No
11	TRPV1	studies	No
10	TRPV1	gene	No

One aspect of causal relationships not yet addressed is whether there are indications of empirical confirmations and how we would go about finding them. While the citing sentences generally do not detail theory-evidence connections, we can learn more about these by looking at neighboring text in the citing papers. For example, looking at the paragraphs containing the citing sentences for the Caterina paper, the word stem "*observ*" ("observe", "observation", etc.) is found an average of four sentences following the citing sentence, suggesting that the observational details are given at a later point in the text. This finding may be important for quantifying the degree of confirmation of theory (or its causal assertion) by evidence, but the nature and strength of this coupling remains to be investigated.

From a broader perspective, I think quantitative science studies can benefit from research into causal networks because it brings us closer to the content and substance of science, scientific theories, theory confirmation, and ultimately the credibility of science.

Acknowledgments

I would like to thank Nees van Eck of CWTS for providing data, Mike Patek for programming, Kevin Boyack and Richard Klavans of SciTech Strategies, and Caleb Smith at the University of Michigan for valuable discussions.

References

Axelrod, R. (1976). The cognitive mapping approach to decision making. In R. Axelrod (Ed.), *Structure of Decision* (pp. 3-17). Princeton: Princeton University Press.

Bird, S., Klein, E. & Loper, E. (2009). *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. Sebastopol, Ca.: O'Reilly Media. http://www.nltk.org/api/nltk.tag.html#module-nltk.tag

Boyack, K.W., Van Eck, N.J., Colavizza, G. & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, 12(1), 59-73.

Bunge, M. (1963). *Causality: the place of the causal principle in modern science.* Cleveland: Meridian Books.

Caterina, M.J., Leffler, A., Malmberg, A.B., Martin, W.J., Trafton, J., Petersen-Zeitz, K.R., Koltzenburg, M., Basbaum, A.I. & Julius, D. (2000). Impaired nociception and pain sensation in mice lacking the capsaicin receptor. *Science*, 288(5464), 306-313.

Findler, N.V. & Bickmore, T. (1996). On the concept of causality and a causal modeling system for scientific and engineering domains, CAMUS. *Applied Artificial Intelligence*, 10(5), 455-487.

Li, Z., Li, Q., Zou, X. & Ren, J. (2021). Causal extraction based on self-attentive BiLSTM-CRF with transferred embeddings. *Neurocomputing*, 423, 207-219.

Narayanan, V.K. & Armstrong, D.J. (Eds.). (2005). *Causal Mapping for Research in Information Technology.* Hershey, Pa: Idea Group Publishing.

Pearl, J. & Mackenzie, D. (1990). *The Book of Why.* New York: Basic Books.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., et al. (2011). Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12,2825–2830.

Small, H. (2020). Past as prologue: Approaches to the study of confirmation in science. *Quantitative Science Studies*, 1(3), 1025-1040.

Sobrino, A., Olivas, J.A. & Puente, C. (2010). Causality and imperfect causality from texts: a frame for causality in social sciences. *International Conference on Fuzzy Systems*. (pp. 1-8) Barcelona: IEEE.

Trieu, H-L., Tran, T.T., Duong, K.N.A., Nguyen, A., Miwa, M. & Ananiadou, S. (2020). DeepEventMine: end-to-end neural nested event extraction from biomedical texts. *Bioinformatics*, 36(19), 4910-4917.

Zeevat, H. & Schmitz, H-C. (Eds.). (2015). *Bayesian Natural Language Semantics and Pragmatics.* Cham, Switzerland: Springer.

Bibliometric indicators to measure the internationalization of Brazilian research: a case study on common beans

Ana Carolina Spatti[1], Luiza Maria Capanema Bezerra[2], Adriana Bin[3], Carlos Eduardo Fredo[4] and Sérgio Augusto Morais Carbonell[5]

[1] *anaspatti@ige.unicamp.br*
University of Campinas, Institute of Geosciences, Rua Carlos Gomes, 250, Campinas, São Paulo (Brazil)

[2] *luiza@iac.sp.gov.br*
Agronomic Institute (IAC), Av. Barão de Itapura, 1481, Botafogo, Campinas, São Paulo (Brazil)

[3] *adribin@unicamp.br*
University of Campinas, School of Applied Science, Rua Pedro Zaccaria, 1300, Limeira, São Paulo (Brazil)

[4] *cfredo@sp.gov.br*
Institute of Agricultural Economics (IEA), Praça Ramos de Azevedo, 254, São Paulo (Brazil)

[5] *sergio.carbonell@sp.gov.br*
Agronomic Institute (IAC), Av. Barão de Itapura, 1481, Botafogo, Campinas, São Paulo (Brazil)

Abstract

This paper aims to measure the internationalization of Brazilian research within the common beans research topic. To achieve this goal, we employed bibliometric techniques using Scopus's bibliographic database. For the analysis, we use diffusion, collaboration and research impact indicators. The results show the predominance of national collaborations and the role of the United States and Brazil in the mainstream of scientific production on common beans. Although there is a strong national appeal on the subject, the analysis of the Field-Weighted Citation Impact indicates that Brazilian publications in international collaboration are more cited than the global average, thus showing the positive impact collaborations have on the global dissemination of science. Furthermore, by offering inputs on the international projection of research, we design a methodological path to assess the internationalization of scientific production by countries and institutions. From this point of view, this paper can support the improvement of research internationalization policies and the strategic planning of science.

Introduction

One of the most relevant criteria to evaluate scientific production is the degree of internationalization (Fiorin, 2007), as it demonstrates the capacity of countries and institutions to produce knowledge for the world scientific community (Santin; Vanz & Stumpf, 2015). In this context, bibliometric studies are gaining ground, which allow visualizing the scientific activity of a given object or field of knowledge, taking into consideration aspects such as collaboration, impact of citations and co-authorship networks (Araújo, 2006; Guedes & Borschiver, 2005).

In this perspective, this article applies bibliometric indicators to measure the internationalization of Brazilian research regarding beans, one of the most produced and consumed legumes in the world, as the object of study (Miyamoto et al., 2017).

Bean is a seed of a leguminous plant of the *Phaseolus* genus (Singh et al., 1991). The species with the greatest representativeness in cultivation and consumption is *Phaseolus vulgaris L.*, also known as "common bean" (Miyamoto et al. 2017, Chiorato et al. 2018). Such legume is an important food and contribute to the food security of several world populations due to the low cost of production, less impact on the family budget and its recognized nutritional quality (Hayat et al. 2014; Messina, 2014; Hartmann & Siegrist 2017).

Methodology

This study applied bibliometric indicators (of scientific diffusion, research collaboration and impact) of papers retrieved from the Scopus platform, due to its scope and volume (Archambault et al., 2009). Scopus is a database of abstracts and citations from scientific literature and academic-level information sources that indexes more than 22,000 journals, 5,000 international editors, in addition to other documents (CAPES, 2016).

From exploratory readings on the theme and support from experts, descriptive terms about the common bean were selected (Table 1):

Table 1. Boolean search equation for common bean

TITLE("Phaseolus vulgaris" OR "common bean" OR "carioca tegument" OR "black tegument" OR "alubia tegument" OR "navy tegument" OR "red tegument" OR "cranberry tegument" OR "pinto tegument" OR "Dark Red Kidney tegument" OR "white tegument" OR "rajado tegument" OR "mulatinho tegument" OR "roxinho tegument" OR "jalo tegument" OR "bolinha tegument" OR "rosinha tegument" OR "bico de ouro tegument")

Note: Search date on November 4, 2020.

Source: Research data.

The search process was carried out using the Boolean formula indicated in Scopus in November 2020, using filtering publications whose terms were present in the "title" field. For capturing the most recent dynamics of scientific production on the subject, we established the 2010-2019 timeframe.

To organize and process data, we used the Scival and VOSviewer tools. Scival is an Elsevier solution based on Scopus' publication records and features metrics on productivity, citation impact and collaboration. VOSViewer, by its turn, is a software to create and explore maps based on network data and bibliometric information.

Findings and Discussion

The application of the search equation in the Scopus database resulted in 3,327 publications (91% of which were articles and the remaining chapters of books, books, reviews and conference articles). Of the total publications, 29% (979) have at least one Brazilian affiliation. Regarding the historical trajectory of publication in the world and Brazil (Figure 1), it is clear that 2018 was the year with the largest number of research products on the theme. In general, there is a global tendency for growth within this research topic, which is also true for Brazil but at a rate almost four times lower.

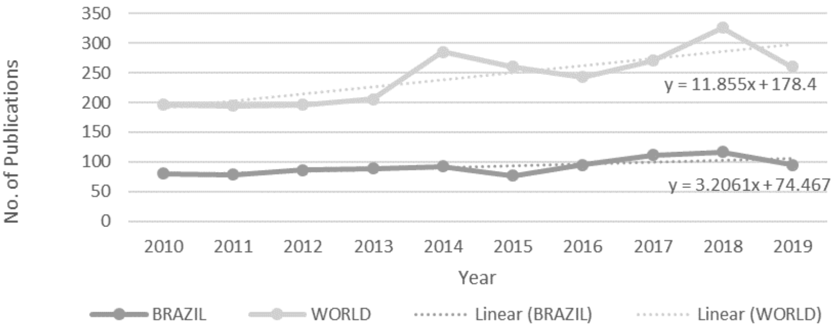


Figure 1. Evolution of publications on common beans over the years (2010-2019)

Source: Research data.

Figure 2 represents the co-authorship network established by the researchers' institutions' home countries. In the figure, the size of the sphere is relative to the number of publications in the country, so that the larger the sphere, the more scientific productions were published by that nation. The connecting lines, by their turn, represent co-authoring relationships. The more intense the connection, measured by the thickness of the line, the greater the number of joint works among researchers from these countries.

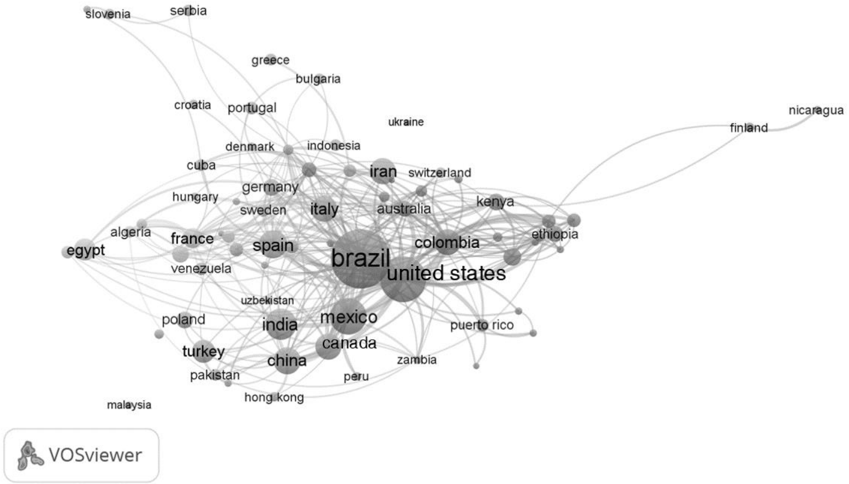


Figure 2. Co-authorship network between countries based on publications on common beans (2010-2019)

Source: Research data.

The network reveals the representativeness, in terms of the number of publications, of Brazil, the United States, Mexico and Colombia and the strong co-authoring relationship between them. Besides, there is the formation of 5 clusters. The first, formed by Brazil, the United States, Mexico, Chile, Colombia and Peru, stands out both for the number of publications and the intensity of the established co-authored relationships. The second (composed of China, India, Turkey, Canada etc.) and third clusters (Italy, Spain, Germany, Indonesia, United Kingdom, Portugal, Japan, etc.) have weaker relationships, but an important number of publications. Egypt, France, Sweden, South Korea, Finland, among others, have moderate indicators of publications and co-authorship.

In Figure 3, the collaboration metrics of the papers are described according to the affiliation to which the authors belong, categorizing them in terms of individual authorship (single), intra-institutional, national collaboration and international collaboration. The data reveal the predominance of national collaborations in publications on common beans, both for the world and Brazil. Such occurrence can be justified by the nature of the research with beans, which can be characterized as applied research and, in general, in line with edapho-climatic conditions and local demands for consumption preference. An example is the case of Brazil where there is a preference for the type of "Carioca" tegument, which is not consumed in other countries.

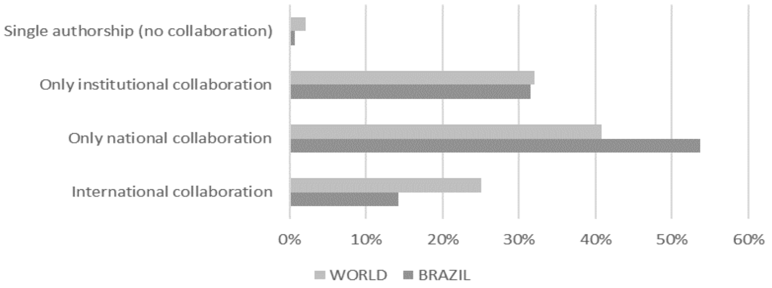


Figure 3. Global collaboration metrics for common beans publications (2010-2019)
Source: Research data.

Regarding the number of countries that cite publications on beans (that is, the countries where the referencing authors' affiliation institutions are located), according to Figure 4, Brazil presents a lower performance than the world. However, within the analyzed period, its publications were referenced, on average, by 80 countries, which indicates considerable global dissemination of Brazil's scientific production on common beans.

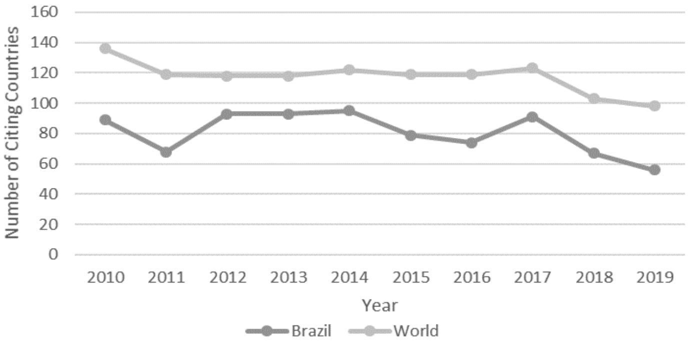


Figure 4. Number of citing countries (2010-2019)
Source: Research data.

Figure 5 presents the collaboration metrics associated with the Field-Weighted Citation Impact (FWCI). This indicator reveals how the number of citations received by a set of publications compares to the average number of citations received by all other similar publications in the data universe. That is, the FWCI reveals how the citations received by publications in Brazil are compared to the world average. An FWCI of 1.00 indicates that Brazilian publications were cited exactly as expected based on the global average of similar publications (with the same period, type and field of knowledge). A FWCI greater than 1.00, in turn, indicates that national publications were cited more than expected based on the global average (ELSEVIER, 2021).

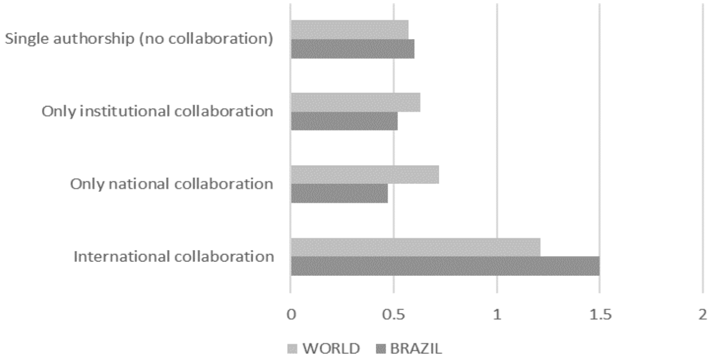


Figure 5. Global collaboration metrics for common bean publications *versus* Field-Weighted Citation Impact (2010-2019)
Source: Research data.

As Figure 5 illustrates, scientific production on common beans is an area of research with relatively low impact, as measured by the FWCI. Only publications resulting from international collaboration reached values greater than 1.00 (which means a positive impact). However, if we analyze the others categories of collaboration (national, institutional and single authorship), the indicator reflects a similar impact behavior in terms of citation of Brazil concerning the world.

Conclusions

The contributions of this paper are twofold. The first refers to the field of Information Science and Science and Technology Management and Planning. By offering inputs on the international projection of research, through indicators of science diffusion, collaboration and impact, we trace a methodological path to assess the internationalization of scientific production by countries and institutions. From this point of view, our paper can aid the improvement of research internationalization policies and the strategic planning of science.

The second contribution is specifically related to studies regarding common beans or studies of similar scope in the context of the Agricultural and Biological Sciences fields. Results show the predominance of national collaborations and the role of countries such as the United States in the mainstream of scientific production on common beans, to which there is a strong national appeal. Moreover, the analysis of the Field-Weighted Citation Impact indicates that Brazilian publications born of international collaboration are more referenced than the global average.

We emphasize that the research topic on common beans can be directly related to the context/environment of ST&I, leading to results of a regional character, which are also influenced by the agricultural technology maturity stage, edaphoclimatic conditions, availability of genetic resources and food security aspects. Thus, it is necessary to take into account the production of social, economic and environmental impacts beyond the scientific dimension. In this sense, we understand that scientific performance indicators based on citations and collaboration networks may underestimate or neglect regional specificities that influence the characteristics of the scientific publication on common beans. Therefore, complementary studies using bibliometric indicators (such as keyword network, institutional link, corresponding author), and other research approaches, can enrich the results presented in this paper.

Acknowledgments

This paper represents a result of the research project "Results and Impacts Evaluation of the Bean Genetic Improvement Program of the Agronomic Institute", supported by the São Paulo Research Foundation (FAPESP), Government of São Paulo State, Brazil – Grant 2020/01721-7, and the National Council for Scientific and Technological Development (CNPq), Brazil – Grant 422680/2018-4.

References

Araújo, C. A. (2006). Bibliometria: evolução história e questões atuais. *Em Questão*, Porto Alegre, v. 12, n. 1, p. 11-32.

Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for information science and technology*, 60(7), 1320-1326.

CAPES (2016). Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Scopus: guia de referência rápida. Available from <https://www.periodicos.capes.gov.br/images/documents/Scopus_Guia%20de%20refer%C3%AAncia%20r%C3%A1pida_10.08.2016.pdf>.

Chiorato, A. F.; Reis, L. L. B.; Bezerra, L. M. C. & Carbonell, S. A. M. (2018). *Global vision on common bean breeding cultivars*. NOVA, New York, p. 27–68. In Phaseolus vulgaris: Cultivars, Production and Uses.

ELSEVIER (2021). What is field weighted citation impact. Available from <https://service.elsevier.com/app/answers/detail/a_id/14894/supporthub/scopus/~/what-is-field-weighted-citation-impact-%28fwci%29%3F/>.

Fiorin, J. L. (2007). Internacionalização da produção científica: a publicação de trabalhos de Ciências Humanas e Sociais em periódicos internacionais. *Revista Brasileira de Pós-Graduação*, 4(8). https://doi.org/10.21713/2358-2332.2007.v4.133

Guedes, V. L. S; Borschiver, S (2005). Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. Encontro Nacional de Ciência da Informação, v. 6, n. 1, p. 18.

Hartmann, C.; Siegrist, M (2017) Consumer perception and behavior regarding sustainable protein consumption: A systematic review. *Trends in Food Science & Technology*, 61: 11–25.

Hayat, I.; Ahmad, A.; Masud, T., Ahmed, A.; Bashir, S. (2014) *Nutritional and health perspectives of beans (Phaseolus vulgaris L.): an overview.* Critical reviews in food science and nutrition 54: 580–592.

Messina, V. (2014) Nutritional and health benefits of dried beans. *The American Journal of Clinical Nutrition,* 100: 437–442.

Miyamoto, B. C. B., Souza, R. F., da Silveira, J. M. F. J., & da Silva Junior, J. J. (2017). Análise da produção científica sobre o mosaico-dourado do feijoeiro. *Revista de Política Agrícola*, 26(3), 79-95.

Santin, D. M., Vanz, S. Na. de S., & Stumpf, I. R. C. (2015). Internacionalização da produção científica em Ciências Biológicas da UFRGS: 2000-2011. *Transinformação*, 27(3), 209-218. https://doi.org/10.1590/0103-37862015000300003

Singh, S. P., Gepts, P., and Debouck, D. G. (1991). Races of common bean (Phaseolus vulgaris, Fabaceae). *Economic Botany,* 45: 379–396.

Differences between Web of Science, Scopus, and Dimensions in structure and citation networks affect German sectors' normalised citation impact.

Stephan Stahlschmidt[1] and Dimity Stephen[2]

[1] *stahlschmidt@dzhw.eu*
German Centre for Higher Education Research and Science Studies, Schützenstr. 6A, 10117, Berlin (Germany)

[2] *stephen@dzhw.eu*
German Centre for Higher Education Research and Science Studies, Schützenstr. 6A, 10117, Berlin (Germany)

Abstract

In 2018 Dimensions was introduced as an alternative bibliometric database to the well-established Web of Science (WoS) and Scopus, however all three databases have fundamental differences in coverage and content, resultant from their owners' philosophies. Considering these differences, we use a citation network analysis and assessment of normalised citation impact of "duplicate" publications to explore whether the three databases offer structurally different perspectives of the bibliometric landscape or if they are essentially homogenous substitutes. Our citation network analysis of core and exclusive 2016-2018 publications revealed a large set of core publications indexed in all three databases that are highly self-referential. In comparison, each database selected a set of exclusive publications that appeared to hold similarly low levels of relevance to the core set and to one another, with slightly more internal communication between exclusive publications in Scopus and Dimensions than WoS. Our comparison of normalised citations for 69,044 publications in all three databases found that German sectors were valuated as more impactful in Scopus and Dimensions compared to WoS, particularly for sectors with an applied focus. We conclude that the databases do present structurally different perspectives, although Scopus and Dimensions vary more from WoS than they do from one another.

Introduction

Just as the introduction of Scopus in 2004 challenged the Web of Science's (WoS) primary position on bibliometric databases, the recent launch of Digital Science's Dimensions may change the bibliometric landscape once more. Like WoS and Scopus, Dimensions has amassed a huge index of scientific documents, however Dimensions has important fundamental differences to its predecessors which may offer a different bibliometric perspective. Further, Digital Science's position as part of the Holtzbrinck Publishing Group, owner of the SpringerNature publishing house, has embedded Dimensions in a content-rich environment that, combined with its relatively open approach, means Dimensions is likely to establish itself in the market as an enduring product. Given the potential uptake of Dimensions for bibliometric studies, it is important we understand how the diverging coverage of Dimensions, WoS, and Scopus might influence the results of bibliometric analyses.

Bibliometric analyses reflect their underlying databases' characteristics in that the databases' coverage fundamentally defines what is included in the analysis and bibliometric evaluation further contextualises the analysed content against the database, again emphasizing its coverage. As such, divergences between databases can have implications for bibliometric assessments. The primary differences between WoS, Scopus, and Dimensions lay in the scope of documents indexed, the content selection processes, the accuracy of internal matching of citing items, and how items are classified to disciplines.

Dimensions indexes content across the full research spectrum, including grants and datasets, beyond the typical content of publications and conference papers prioritized in Scopus and WoS, and as such, it is a much larger database. These differences result from the database owners' varying philosophies: WoS and Scopus use selection panels to curate their content to that of a certain quality – with WoS enacting a higher threshold than Scopus – whereas Digital Science applies no editorial judgement to what is indexed in Dimensions. Within the databases, there are differing levels of accuracy in matching documents that cite one another. Visser, van Eck, and Waltman (2020) found issues with missing reference lists and subsequent incomplete citation links in Dimensions, while missing references, incorrect reference metadata, and incorrect item matching are a problem in WoS and Scopus (van Eck and Waltman, 2017), although less extensively so than in Dimensions. Also, WoS and Scopus classify items to disciplines based on the subject orientation of the publishing journal, while Dimensions classifies individual items based on their content. These differences all have important implications for the set of documents against which a publication is normalized and compared in the context of each database.

As Dimensions was only recently launched, so far only a small number of studies have examined the differences between it, WoS, and Scopus that result from these differing practices and philosophies. Orduña-Malea & Delgado-López-Cózar (2018) compared samples of Library and Information Science documents, authors, and journals in Scopus and Dimensions and found that coverage in Dimensions exceeded that of Scopus, but Dimensions recorded fewer citations. Thelwall (2018) tested Dimensions' retrieval of nearly 90,000 food science articles with DOIs in Scopus, alongside a random sample of 10,000 articles from 2012. Over 90% of the food science articles were captured in Dimensions, and the citation counts in both databases were highly correlated (0.9-1.0), leading Thelwall to conclude that Scopus and Dimensions were interchangeable on coverage and citations. Harzing (2019) compared Dimensions, Scopus, and WoS on retrieval of her own publication corpus and six key Business and Economics journals. She found Dimensions and Scopus were approximately equal in both their coverage and citation counts, and both produced higher measures than in WoS.

Visser et al. (2020) compared WoS and Dimensions to Scopus, identifying Dimensions as the largest database, although there was substantial overlap in content between Scopus and both WoS (overlap of 17.7 million documents) and Dimensions (21.3 million). However, the share of Dimensions' content not in Scopus was nearly double (40.9%, 14.8 million) that of WoS (22.7%, 5.2 million). Martín-Martín, et al. (2020) also noted that, while Scopus and Dimensions offered twice as much exclusive content as WoS, the databases contained a high degree of overlapping content; 75-78% of their sample overlapped in pair-wise comparisons and 66% was present in all three databases. The largest content convergence occurred in the hard sciences, where 73-78% was in all three databases, with more divergent content in the medical sciences and engineering (65% in all three databases), social sciences (54%), and humanities (41%).

Coverage differences between WoS and Scopus have already been shown to influence the outcome of bibliometric analyses (Stahlschmidt & Stephen, 2019; Huang et al., 2020), however seemingly no study has yet examined such differences between WoS, Scopus, and Dimensions. As such, here we use a quantitative comparison to determine the extent to which bibliometric indicators are influenced by the characteristics of the three databases. We do this by first analysing the databases' citation networks. Previous studies show that the divergent coverage between the databases results in different subsets of publications indexed in only one, two or all three databases (Visser et al., 2020; Martín-Martín, et al., 2020). The power set of these subsets consists of the set of publications and citation links between them that are contained in all three databases, partial intersections of publications and related citations in two databases, and the residual sets of exclusive publications and related citations indexed in only one database. We investigate the role of these subsets for themselves and other subsets. With respect to citation analyses, role is understood as relevance and impact, which can be measured by reducing citations to their Mertonian ideal ("give credit, where credit is due"), by citations between and within these subsets. Citations between subsets represent how information flows and how the publications indexed exclusively in a database are embedded in this information flow. The role of these publications in the overarching citation network, which is unknown in its population, can be identified and the added value of each database as a reflection of previously unobserved scientific communication can be approximated.

Using the context provided by the citation network, we then assess the differences in normalized citation impact of duplicate publications between the three databases. In bibliometrics, priority is given to normalised indicators that evaluate a publication in relation to its environment. Due to the different coverage of the databases, environment-specific differences arise in the evaluation of the same publication. We therefore analyse the same publications in the different databases and determine how each publication's valuation changes given the environment of the databases against which it is normalised. This varying evaluation of the same content can be used to illustrate the structural differences in the databases, which is informative for interpreting bibliometric analyses (Stahlschmidt & Stephen, 2019). Through these two analyses, we examine whether the databases offer a slightly varying but essentially homogeneous representation of the general citation network and are therefore substitutes, or whether the databases show structurally different perspectives.

Methods

We sourced the data for our analyses from the German Competence Centre of Bibliometrics' (KB) in-house versions of WoS and Scopus databases. For WoS, this version includes the Science Citation Index Expanded, Social Science Citation Index, and Arts and Humanities Citation Index. Scopus and Dimensions databases are not organised into indices and as such, we used the relevant documents from the entire databases. Dimensions data are a snapshot of the database as of September 2019 and WoS and Scopus data are snapshots as of April 2019.

Citation network analysis

We analysed articles and reviews published in 2016-2018 and citations in a three year window between 2016 and 2018. We restricted our analysis to articles and reviews, however a known issue in Dimensions is that all documents in journals are assigned document type “article” (Visser et al., 2020). For instance, Dimensions holds more documents of type “article” in 2016-2018 than the intersection of core publications jointly indexed by WOS, Scopus and Dimensions does. However, most of these documents do not include any references. In comparison, only 1% of WOS and 4% of Scopus articles and reviews in 2016-2018 had no source references, and no WoS and only 9 Scopus articles had no references at all. To address this issue, we selected articles and reviews indexed in WoS and Scopus, and articles in Dimensions with at least one reference. Thus for Dimensions we separated the substantial scientific contributions that build on and highlight former contributions via references from other journal content to improve the validity of the database comparison. Still this requirement of at least one reference for Dimensions articles constitutes a lower bound rather than a solution to the insufficient document type classification in Dimensions, but WOS and Scopus also disagree sometimes upon the document type.

We joined the three databases via an exact string matching procedure based on DOI identifiers. DOIs uniquely identify publications and are therefore suitable for matching purposes. DOIs recorded in bibliometric databases have partially been observed to include errors (Akbaritabar & Stahlschmidt, 2019), e.g. non-unique DOIs or misread string characters in optical character recognition processes, however these issues might arise randomly and might not adversely affect any structural differences between databases. Instead DOI matching has been observed to produce highly valid matching results (Fraser & Hobert, 2019) and less than 7%, 10%, respectively 1% of articles and reviews in 2016-2018 indexed in WoS, Scopus or Dimensions were missing a DOI.

The citation network analysis is implemented by the indicators (1) out/in-degree, as a network analytical perspective on the micro level of publications, which is translated to the level of the subsets previously described by representing the respective distributions, and (2), as far as possible the indicator internal coverage, as an aggregated value on the database level. With this approach, in addition to the expected increased coverage in Dimensions due to its larger size, i.e. the pure "more" of communication, communicative characteristics of the databases can also be partially quantified, which provides useful context for the normalised citation analysis.

Normalised citation analysis

We selected German sectors as the level at which we would assess the effect of database choice on normalised citation impact. We first identified all articles published in 2016 affiliated with German institutions that were indexed in all three databases, which we refer to as "duplicate" publications. We retrieved the WoS-Scopus duplicates, which were identified by comparing hash values on a subset of metadata strings between the two databases. We then extracted all German articles published in 2016 from Dimensions and matched their DOIs to the DOIs from Scopus of the WoS-Scopus duplicates to identify the documents in all three databases. We then validated the DOI-based matches by calculating the Jaro-Winkler distance on the title strings between the three versions using *stringdist* in R (van der Loo, 2014).

To assess the effect of the database's environment on the normalised citations, we calculated for every German duplicate the number of citations the article received between 2016 and 2018 from each database (observed citations), and the average number of citations received in these three years by all articles published in 2016 that were allocated to the same discipline (expected citations). We then calculated the difference Δ in normalised citations between databases as:

$$\Delta\, norm.cit. = \frac{obtained\ citations_i^{(s1)}}{expected\ citations_i^{(s1)}} - \frac{obtained\ citations_i^{(s2)}}{expected\ citations_i^{(s2)}}$$

where *i* is each German duplicate, and *s* is the source database. We normalised each duplicate's citations against documents of the same type and discipline as they were defined in each database. This necessitated that we exclude unclassified documents. As we wanted to examine the effect of using each data source in its "natural" state as much as possible, we did not control for the differences between databases in the disciplines to which publications are assigned. Further, as each database allows documents to be assigned to multiple disciplines, there were often multiple contrasts made for a single duplicate between databases for each combination of disciplines. To account for this, we calculated the difference in normalised citations for each combination of disciplines for each duplicate, then averaged the differences to obtain one summary observation of the differences in normalised citations of the duplicate per database combination. We then aggregated the duplicates to sectors on a whole-counting basis.

In normalising the observed citations in each database against the expected citations in the same database, we achieved a database-specific valuation of each article. The content of the database is influential here as the inclusion or exclusion of particular articles in the corpus may influence both the citations received by the article and the average citations received by all articles in the discipline, affecting the ratio between the two observations. In examining the difference in normalised citation impact between databases, we can examine how the same publication can be valuated differently between databases due to the database's environment.

Results

Citation network analysis

We show in Figure 1 the number of articles and reviews published in 2016-2018 in each database, and the intersection and exclusive content in each combination of databases as identified through the matching process. Dimensions indexed the largest number of publications (>6.3 million), followed closely by Scopus (5.9 million), while WoS indexed substantially fewer documents (4.8 million). The difference between Dimensions and Scopus may actually be smaller than reported here as the aforementioned imperfect lower bound that Dimensions articles contained at least one reference might still allow other non-article documents to be included. The majority of publications (4.3 million) were indexed in all three databases and this set of core publications is by far the largest intersection (67%, 71%, 88% of the entire Dimensions, Scopus, and WoS corpuses, respectively). Dimensions indexed 1.3 million exclusive publications, or about 20% of its entire corpus, and Scopus exclusively indexed 0.6 million publications or ~10% of its corpus, while WoS' exclusive publications constituted less than 1% of its corpus. Hence WoS differentiated itself from other databases not by exclusively indexed publications, but seemingly by foregoing the indexation of more publications. Notably, contrary to its inclusive indexing policy, Dimensions apparently does not index some 1.05 million documents indexed in either or both WoS and Scopus.

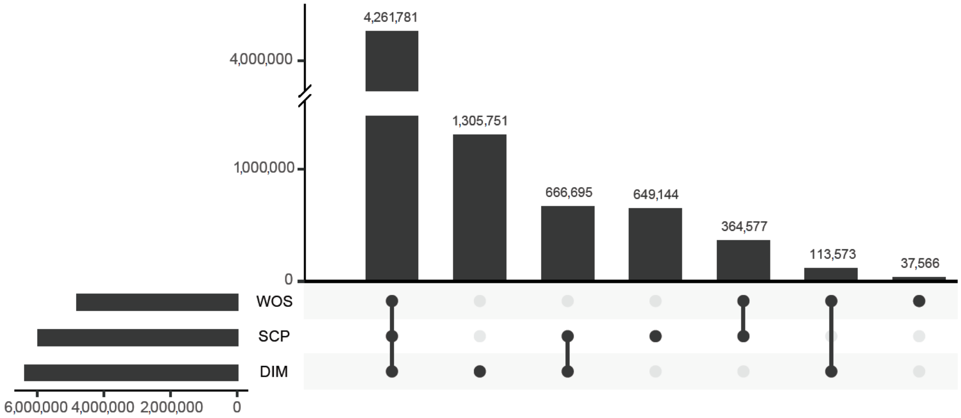


Figure 1. Number of publications with a DOI in 2016-2018 (left panel), magnitude of intersections (upper panel) and relation between total databases and intersections (lower panel).

Given the sizable differences between the databases, we analysed how the internal coverage of core publications varied due to the different indexation practices. By observing if a reference in an indexed publication is or is not itself indexed in the same database we compared the relevance attributed to the work by the author with the relevance attributed by the database provider. A pronounced difference in assumed relevance, manifesting as low coverage, indicates that a database only partially captures the communication flow perceived relevant by authors and hence a bibliometric analysis might be biased on any such out-of-sync dataset.

As documents in the core set of jointly indexed publications have unanimously been deemed relevant by the three database providers and hence allow for a comparison, we used their reference lists to observe potential differences in database-specific coverage. References to non-core publications in particular differentiate the databases in their coverage. Figure 2 shows the database-specific percentage of indexed, or so-called source references, via a density plot across the 4.2 million core publications. The upper panel shows the internal coverage of all 2016-2018 publications, while the lower panel depicts the share of indexed references published in 2016.

The overall internal coverage shows large variability with values from zero to one hundred percent. All three distributions are skewed to the left indicating that although a large share of references in core publications were indexed, some indexed core publications had few of their references indexed in the respective database. In particular, the social sciences and humanities, with their reduced focus on journal articles as the primary communication device, have been observed to lack internal coverage (Stephen, Stahlschmidt & Hinze, 2020). Apart from this discipline-specific effect affecting all databases, we observed notable differences in the overall internal coverage between databases. Whereas Dimensions and WoS demonstrated relatively high internal coverage, with over 90% of references from many core publications also indexed, Scopus had lower agreement with the authors' relevance attribution.

The lower panel shows the three-year citation window perspective as we restricted the analysis to references from 2016-2018 publications to 2016 publications. We observed no pronounced visual difference between the databases here, however on average only approximately 5% of all references are considered and most other signals of relevance attribution via references are discarded. As the scale of the x axis might conceal actual differences between the databases, in the following analyses we focused especially on this lower end of 2016 publications. In doing so we adopted a maximum contrast approach comparing core publications indexed in all three databases with exclusive publications indexed solely in one of the three databases.

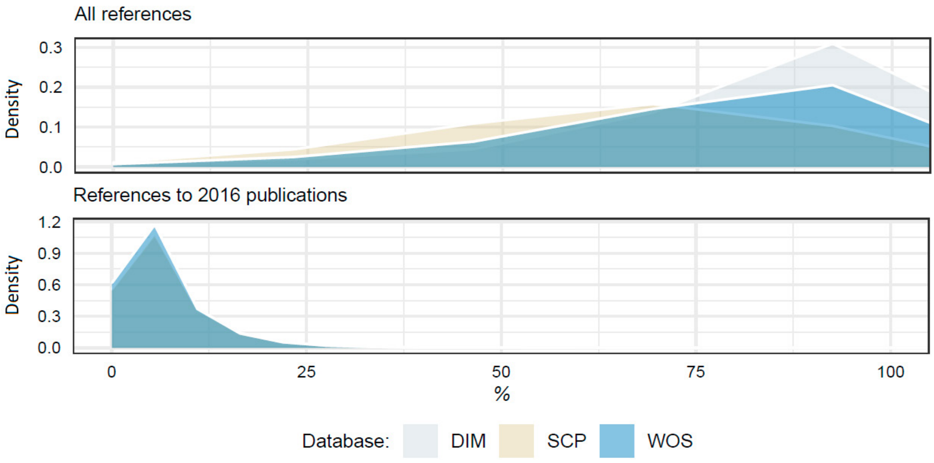


Figure 2. Internal coverage of references in jointly indexed publications published 2016-2018.

We commenced the analysis of references to 2016 publications by examining the internal communication patterns of core publications. Figure 3 shows the share of 2016-2018 core publications that cited a core 2016 publication. As to be expected by the overall low share of references to 2016 publications depicted in the lower panel of Figure 2, the share of internal communication is rather small and the majority of 2016-2018 publications did not reference a 2016 publication. This observation holds for all databases and hence they can only be compared in the following analyses on the residual share of publications that did reference a 2016 publication. We observed that core publications possessed a high degree of internal communication with a sizeable number of core publications referencing other core publications. Hence these publications define a highly self-referential, interlinked network component. This component was identified and indexed by all three databases and constituted between 67% (Dimensions) and 88% (WoS) of the databases.

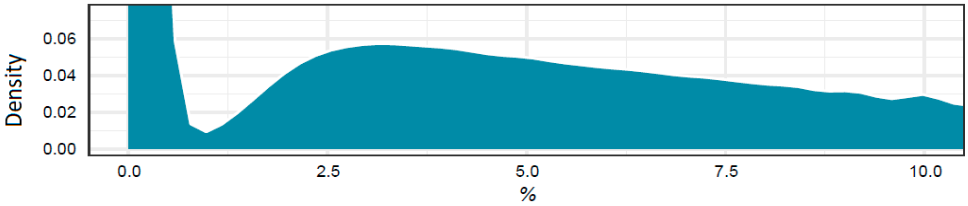


Figure 3. The share of internal references from core 2016-2018 publications to core 2016 publications.

We then examined how core and exclusive publications are interlinked. In Figure 4 we show the share of references in exclusive 2016-2018 publications in each database to core 2016 publications. As Digital Science currently only provides source references for publications indexed in Dimensions, the total number of references in Dimensions-exclusive publications is unknown and for comparability we report instead the share of 2016 publications among all source references of 2016-2018 publications. Comparing the three databases, especially publications exclusive to WoS exhibited a strong dependence on core publications, resulting in a denser overall database than Scopus and Dimensions. But exclusive publications in these two larger databases also rely to a large and similar extent on core publications. Considering the different number of exclusive publications (Figure 1), WoS seemingly foregoes indexing more publications, but offsets this with a more dense citation graph.

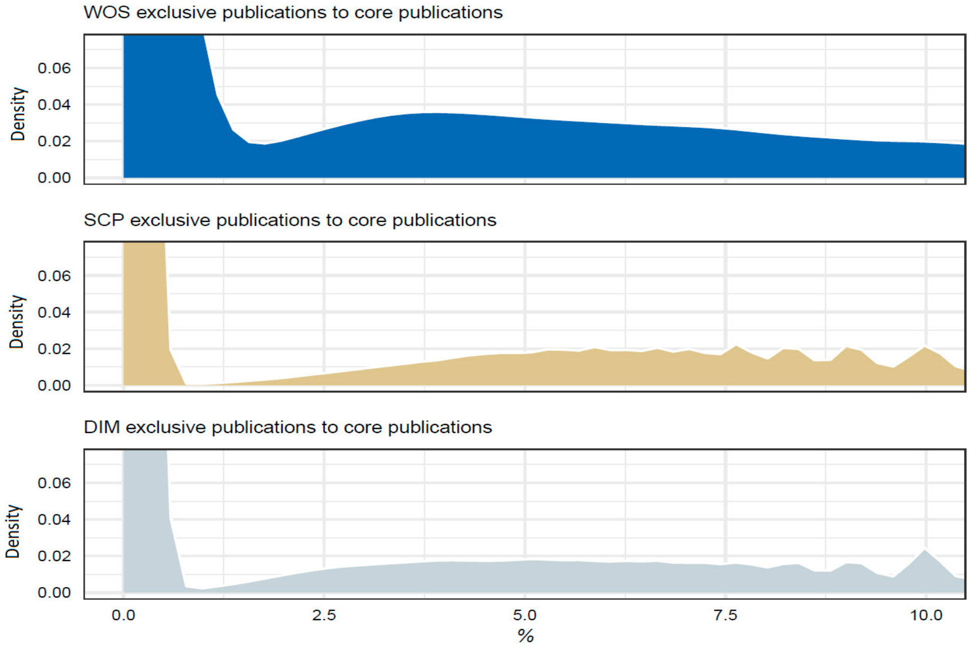


Figure 4. The share of references in exclusive 2016-2018 publications to core 2016 publications.

To complete the interlinkage between core and exclusive publications, we depict in Figure 5 the share of references in core 2016-2018 publications citing exclusive 2016 publications. As we could observe the total number of references of these core publications via WoS and Scopus, we normalised by the total reference count and not the source reference count used before when the total number of references in the Dimensions-exclusive publications was unknown.

A notable share of core references linked out to exclusive publications across all databases. However, considering the actual percentage is often less than 2.5%, it appears that the exclusive publications are of low relevance to core publications. In comparison, the range for internal references among core publications in Figure 3 and the share of core publications referenced in exclusive publications in Figure 4 is much higher, often above 5%. Hence, the content of core publications seemed to carry more relevance than the content of exclusive publications. Exclusive publications might then be understood as an outer, consecutive circle building upon the content provided in core publications. The lower density in WoS compared to Scopus and Dimensions may reflect that its 38,000 exclusive publications may not appear in the 4.2 million core publications as often as Dimensions' 1.3 million exclusive publications.

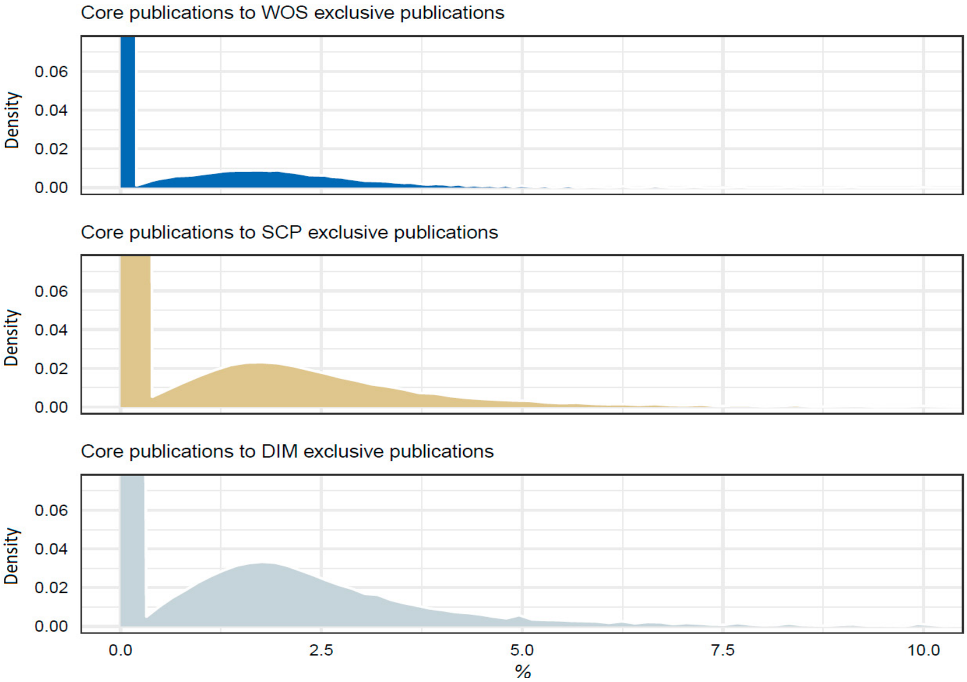


Figure 5. The share of references in core 2016-2018 publications to exclusive 2016 publications.

Normalised citation analysis

For the citation analysis, we matched 2016 publications affiliated with a German institution across databases. The KB WoS-Scopus matching process identified 107,800 of the 113,227 WoS publications (95.2%) and 127,542 Scopus publications (84.5%) as duplicates. We identified 118,688 German publications in Dimensions, however this is an under-estimate as up to 43% of Dimensions records were missing information about the publishing country in the applied Dimensions 2019 snapshot. Via DOI matching we identified 84,332 publications that were in all 3 databases, which was 74.5% of the total 2016 German publications in WoS, 71.1% in Dimensions, and 66.1% in Scopus. After removing review documents, and records missing discipline or sector data, we had a final sample size of 69,044 duplicates. We calculated the Jaro-Winkler distance on the combinations of the duplicates' titles to validate the DOI matching. The average distance was 0.02 and we manually confirmed cases above 0.25, concluding that the DOI-based matching was accurate.

As a preliminary examination of the differences in citations between the databases, we present in Figure 6 pair-wise comparisons of the citations observed for each duplicate article between databases. The diagonal lines represent a perfection correlation. Positive correlations were evident between each database, suggesting all three databases present a comparable picture of the bibliometric landscape, however there appeared to be a slight trend toward higher citations in Scopus and Dimensions compared to WoS. The greatest variation was between WoS and Scopus (mean difference = 1.2), with slightly less between Dimensions and WoS (1.0), and least between Dimensions and Scopus (0.2). These variations from a linear trend due to the databases' exclusive content generate changes in the ratio of observed to expected citations, which translates to variations in publications' normalised citation impact between databases.

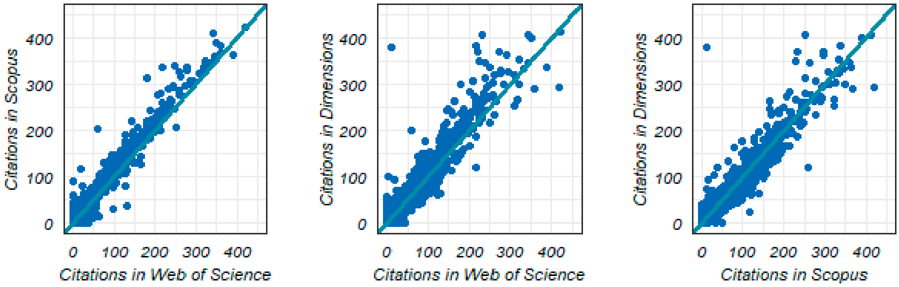


Figure 6. Pair-wise comparisons of duplicate articles' observed citations in each database

We see in Figure 7 the macro effect of the databases' structural differences via the differences in normalised citation impact of the six German sectors. Data are presented as the density in the distribution of differences in each sector's 2016 duplicate publications, and divided into quintiles. The sectors are the Leibniz Association (WGL), the Max Planck Society (MPG), the Helmholtz Association (HGF), the Fraunhofer Society (FhG), universities and colleges of applied science constitute the higher education institutions (HEI), and the business sector (Economy). Each sector has a particular research profile. The universities undertake both teaching and research in all disciplines, while the colleges focus on technical application in specific areas. The HGF has a health, energy, earth and physical sciences orientation focusing on research infrastructure. The WGL conducts basic and applied research in engineering and social, health, and natural sciences, the MPG conducts primarily basic research, and the FhG focuses on applied research.

We see in the bottom panel of Figure 7 that the majority of publications in each sector had higher normalised impact valuations in Scopus compared to WoS. The FhG and Economy sectors benefitted most strongly from Scopus's exclusive content, with around 80% of these sectors' publications improving in impact, compared to around 70% in the other sectors.

Dimensions' content and characteristics also improved the impact of all sectors compared to WoS, as shown in the middle panel, although the effect is not as strong as between WoS and Scopus. The increased impact is nearly uniform across sectors, with the 40% of publications constituting the central and middle-high quintiles in each sector increasing in impact by up to 25%. The MPG lost slightly more impact in Dimensions than the other sectors, particularly when compared to the FhG. Finally, the differences in impact between Scopus and Dimensions in the top panel are more normally distributed, however there is a slight skew toward improved impact in Dimensions, particularly for the FhG where nearly 60% of publications improved.

Overall then, the larger content and specific characteristics of Scopus and Dimensions appears to produce higher normalised citation scores than in WoS, particularly for the sectors with a focus on applied sciences, such as the FhG and Economy sectors. However, the differences in content between Dimensions and Scopus produced less notable differences in normalised impact at this macro level.

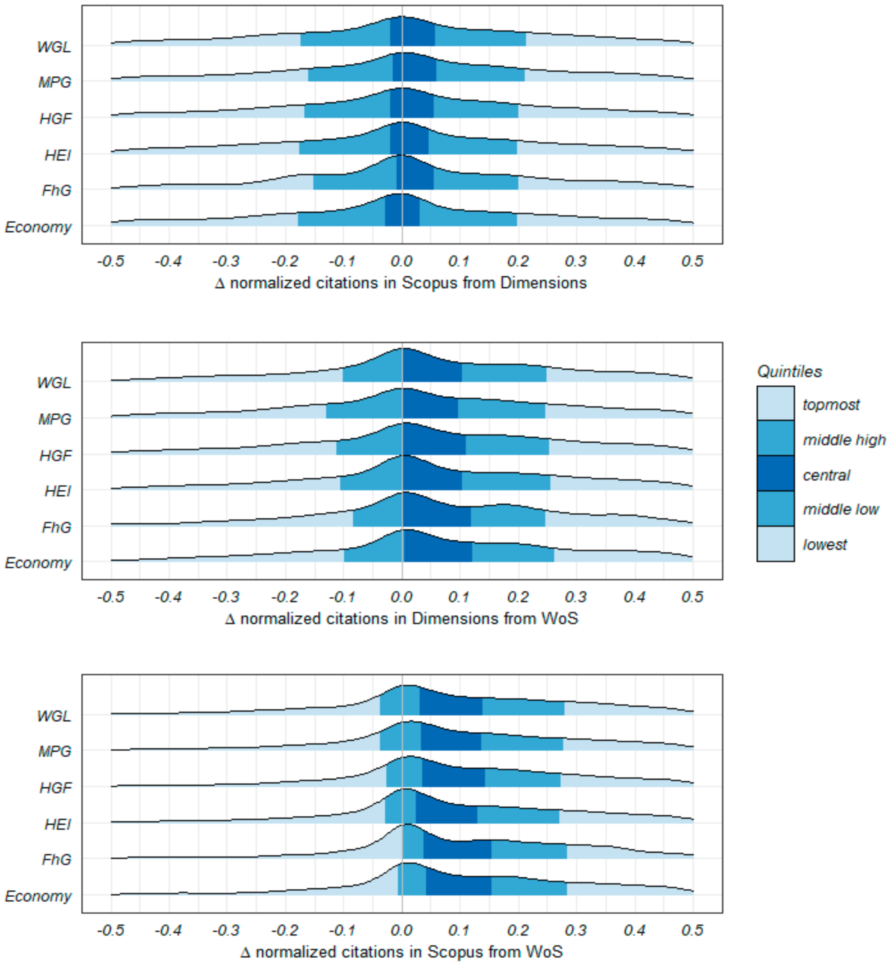


Figure 7. Distribution and quintile of differences in duplicates' normalized citations between databases by German sector.

Discussion

WoS, Scopus, and Dimensions databases differ in particular fundamental characteristics, such as the number and type of documents indexed, the completeness of citation links between documents, the inclusion criteria applied to content, and how items are classified to disciplines. Previous studies have found these differences resulted in variation in database size and coverage, although there was substantial overlap in the databases' content, and that the databases produced similar citation counts (Martín-Martín et al., 2020; Visser et al., 2020; Orduña-Malea & Delgado-López-Cózar, 2018; Thelwall, 2018; Harzing, 2019). However, no study had yet compared the three databases on normalised bibliometric indicators. In this study, we analysed the citation networks and differences in normalised citation impact between the databases to investigate whether the exclusive content of the databases meant they offered structurally different perspectives of the bibliometric landscape.

In our citation network analysis, we identified core publications jointly indexed in all three databases and exclusive publications solely indexed in each of the three databases. In a maximum contrast approach we compared the communication flows via references across these sets to observe the analytical value each database revealed by its particular indexation practices. Not unexpectedly, the indexation of more publications was accompanied by an offset in the density of communication flows in the resulting citation graph, particularly in Dimensions. In all databases, the group of core publications might be characterised by its relatively high degree of self-reference, where former core publications constitute the base or frame for new core publications. Exclusive publications appeared to mark an outer ring drawing substantially upon core publications, hence knowledge is transferred from the core to the outer circle, and this transfer was especially visible in Scopus and Dimensions. However, exclusive publications in all databases exhibited low relevance to the core, denoted by low citation of exclusive publications by the core. The emerging model of knowledge transfer within databases appears to be that of a star, wherein an interconnected core produces knowledge in a self-referential mode, which is then disseminated to the outer circle. Notably, this model persisted across databases because the particular choice of exclusive publications by a database resulted in no visual difference in the citation of these publications by the core, signaling that no database apparently managed to identify particularly relevant publications for the core publications. Instead, all three databases found different exclusive publications of the same low relevance to the core. As a seeming paradox, exclusive publications in this sense do not distinguish databases from one another, but constitute different samples with the same characteristics. As a consequence, core publications might not be enhanced by any of the three perspectives offered by the databases.

The combined effect of larger exclusive content, similar reliance on core publications, but more internal communication between exclusive content in Scopus and Dimensions than WoS (unreported) appears to increase the normalised citation impact of all German sectors in Scopus and Dimensions compared to WoS, in particular for those sectors with an applied focus. The applied science sectors, such as the Economy sector, likely especially benefit due to the better coverage of the social sciences in Scopus than WoS (Stephen et al., 2020). Scopus and Dimensions' similar levels of communication between and within the core and their choice of exclusive publications means there was little difference in impact between the databases.

We caution that, given the relatively small difference between Scopus and Dimensions in absolute article counts, the imperfect document type resolution in Dimensions, the likelihood that the article document type in Dimensions has not been perfectly singled out by our reference requirement but is actually over-reported, the additional 6 months of content in the Dimensions data, and the reliance on only a small portion of references as the relevance attribution, a sizeable amount of incertitude accompanies the current study.

In summary, we conclude that the databases' structural differences mean they provide different bibliometric perspectives, and that Scopus and Dimensions vary more from WoS than they do from one another. WoS with its restrictive indexation policy and Scopus with its selective indexation policy might constitute two separate self-imposed stances with a distinct message: WoS largely represents the well-interconnected core citation network component, while Scopus allows us to observe some transfer from the core to the periphery, resulting in increased normalised impact for most core publications. Dimensions with its laissez-faire indexation policy conveys more coverage but a less decisive, although similar, message to Scopus.

Acknowledgments

Data in this study were obtained from the German Competence Centre for Bibliometrics, funded by the Federal Ministry for Education and Research (BMBF) with grant number 01PQ17001.

References

Akbaritabar, A. & Stahlschmidt, S. (2019, September). Merits and limits: Applying open data to monitor Open Access publications in bibliometric databases. In G. Catalano, C. Daraio, M. Gregori, H. F. Moed, & G. Ruocco (Eds.), *Proceedings of the 17th International Conference on Scientometrics and Informetrics (Vol. 2)* (pp. 1455-1461). Rome: Edizioni Efesto.

Fraser N. & Hobert, A. (2019). Report on Matching of Unpaywall and Web of Science. Tech. rep.

Harzing, A. W. (2019). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics, 120*(1), 341–349. DOI: 10.1007/s11192-019-03114-y.

Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K. & Ozaygen, A. (2020). Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies, 1*(2), 445–478. DOI: 10.1162/qss_a_00031.

Martín-Martín, A., Thelwall, M., Orduna-Malea, E. & Delgado López-Cózar, E. (2020). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, DOI: 10.1007/s11192-020-03690-4.

Orduña-Malea, E. & Delgado-López-Cózar, E. (2018). Dimensions: Re-discovering the ecosystem of scientific information. *El Profesional de la Información, 27*(2), 420-431. DOI: 10.3145/epi.2018.mar.21.

Stahlschmidt, S. & Stephen, D. (2019, September). Varying resonance chambers: A comparison of citation-based valuations of duplicated publications in Web of Science and Scopus. In G. Catalano, C. Daraio, M. Gregori, H. F. Moed, & G. Ruocco (Eds.), *Proceedings of the 17th International Conference on Scientometrics and Informetrics (Vol. 2)* (pp. 1698-1709). Rome: Edizioni Efesto.

Stephen, D., Stahlschmidt, S. & Hinze, S. (2020). Performance and Structures of the German Science System 2020. Studien zum deutschen Innovationssystem. Studie 5-2020. Berlin: EFI.

Thewall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? *Journal of Informetrics, 12*(2), 430-435. DOI: 10.1016/j.joi.2018.03.006.

Van der Loo, M. (2014). The stringdist package for approximate string matching. *The R Journal*, 6, 111-122, https://CRAN.R-project.org/package=stringdist.

Van Eck, N. J. & Waltman, L. (2017, September). Accuracy of citation data in Web of Science and Scopus. In R. Rousseau, W. Glänzel, & Z. Rongying (Eds.), *Proceedings of the 16th International Conference on Scientometrics and Informetrics* (pp. 1087-1092). Wuhan: HSE.

Visser, M., van Eck, N. J. & Waltman, L. (2020). *Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic*. Retrieved December 11, 2020 from: https://arxiv.org/abs/2005.10732.

PubMed Central Citation Context Dataset

Fengjun Sun[1], Yingqiu Li[1], Guojun Sheng[1] and Xiaolin Yao[1]

[1] *{sunfengjun, liyingqiu, shengguojun, yaoxiaolin}@neusoft.edu.cn*
College of Information and Business Management, Dalian Neusoft University of Information, Dalian 116023, (China)

Abstract

The last few decades have witnessed the rapid growth of scientific publications and scholar data, and citation context is one of the most frequently used data in scientometrics and natural language processing. However, current citation context datasets have limitations in several regards. Therefore, we propose a large citation context dataset based on the open accessed PubMed Central papers. In total, we generated about 97.5 million citing-cited document pairs, 34.7 million paragraphs of citation context based on 2,658,541 PubMed Central papers of biomedical and life sciences. The dataset can not only be applied for citation analysis, but also can be used for biomedical text mining and other natural language processing tasks.

Introduction

Citation context is the text around the citation marker in the body text of a citing document. Even though the citation context is embedded in the citing document, it is also a short description about the citing document from the citing authors' point of view. Some previous studies treated citation contexts as an alternative representation of the cited document (Bradshaw, 2002, 2003; Ritchie, 2008). The applications of citation context include summarization, synonym identification and disambiguation, entity recognition and relation extraction, curation and information retrieval, and citation context can also be a source of unannotated comparable corpora (Nakov, Schwartz, & Hearst, 2004). Though early studies of citation context were conducted in the late 1970s (Small, 1978) and early 1980s (O'Connor, 1982, 1983), there are not many studies in this area until the late 1990s mainly due to availability of digital scientific documents. After the upcoming of the Internet era, more and more machine-readable scientific literatures are available online (Lawrence, Giles, & Bollacker, 1999), and studies on citation context have gained much interest among the scientific community in the last two decades. A few reviews about the studies of citation context can be found in the last few years (Bornmann & Daniel, 2008; Ding et al., 2014; Jha, Jbara, Qazvinian, & Radev, 2017; Zhang, Ding, & Milojević, 2013).

However, there are only a few publicly available large citation context datasets for research community, and existing datasets have limitations in different regards. A recent study (Saier & Färber, 2020) compared several scholarly dataset, which include CiteSeerX (Huang, Wu, Liang, Mitra, & Giles, 2015), Scholarly Dataset 2 (Sugiyama & Kan, 2015), arXiv CS (Färber, Thiemann, & Jatowt, 2018), ACL-ARC (Bird et al., 2008), ACL-AAN (Radev, Muthukrishnan, Qazvinian, & Abu-Jbara, 2013), and PubMed Central Open Access Subset. These datasets might contain noisy, inaccurate information, and some datasets are in PDF format, which is hard to process for computers even with the modern artificial intelligence technology.

In this paper, we aim to give a description of a large citation context dataset (Sun, 2020), specifically, the citation context dataset extracted from PubMed Central articles. PubMed Central (PMC) is the online repository for free full text of biomedical and life sciences published journal papers (Roberts, 2002). It is operated by the National Center for Biotechnology Information, a division of the National Library of Medicine at the U.S. National Institutes of Health. Until March of 2021, there are about 6.9 million articles in the PMC archive. However, even though the full text of the PMC papers is freely accessible online, some of the articles cannot be downloaded in bulk due to the operator's policy, and only a few collections of the PMC files are permitted for bulk retrieval. Theses collections are provided for the task of text mining for researchers, which are Open Access Subset, Author Manuscript Collection, and Historical OCR Collection. Among the three collections, the Open Access Subset (OA Subset) we used in this paper is the largest collection available for text mining via PMC and the subset has been used in several recent studies (Boyack, van Eck, Colavizza, & Waltman, 2018; He & Chen, 2018; Small, 2018; Small, Tseng, & Patek, 2017). Articles in this subset are still protected by copyright laws, however, they are made available for bulk download under license such as Creative Commons, which give users the liberty for redistribution and reuse about the data. The files of the articles in the collection are in three formats (PDF, XML, TXT), and we chose to process the XML files, because XML format is designed to be both human- and machine-readable and has tags to locate specific elements in the file, thus simplifying our later processing of the full text of the PMC paper.

Even though, the full text of PubMed Central articles can be a great resource for citation-based tasks with clean, heterogeneous annotation of citations, however, as the study (Gipp, Meuschke, & Lipinski, 2015) pointed out, the collection of full texts don't provide relationship between documents, and the mixed usage of identifiers is not convenient for users to gain accurate citing-cited relationships between documents. On the one hand, we might utilize the XML annotations of document identifiers to reveal the explicit citing-cited relationships between documents. On the other hand, the full text mixed with XML tags itself provides noise for information tasks such as natural language processing. Thus, we need to eliminate the XML tags, leaving clean and pure text of citation context to be processed for future users.

Methods

As our data is based on the PMC text mining open dataset, we first downloaded the dataset in late December of 2019 using the official FTP service. The downloaded PMC XML files are in eight compressed files with the file extension "tar.gz", in which four of compressed files can be used commercially and the other four compressed files can only be used non-commercially. First, we uncompressed all 8 compressed files and saved the PMC files in one directory. There are 2,658,541 PMC papers in total for this dataset. Each PMC file is in the XML format with the file extension "nxml".

The source code of the program to generate the citation context dataset from the original PMC XML files can be found on github: < https://github.com/ku69vOZ8/PubMed-Central-Citation-Context>. To build the PMC citation context database, we need to process each of the PMC paper and get the necessary information within the PMC paper's XML file, including PubMed ID (or PMID, if available), local path, cited documents, citation context of each of the cited documents. The program to process the PMC files includes the following steps (see Figure 1):

(1) extract the PMC paper's metadata.
(2) extract the citation relationships.
(3) extract the citation contexts.

As it is shown in Figure 1, we can obtain the PMC paper's metadata from the full text for step 1, citation relationship from reference list for step 2, and citation contexts based on both full text and reference list for step 3. We use 4 entities in our program and database, which are Literature, Cite, CitationContext and CitationContextText.

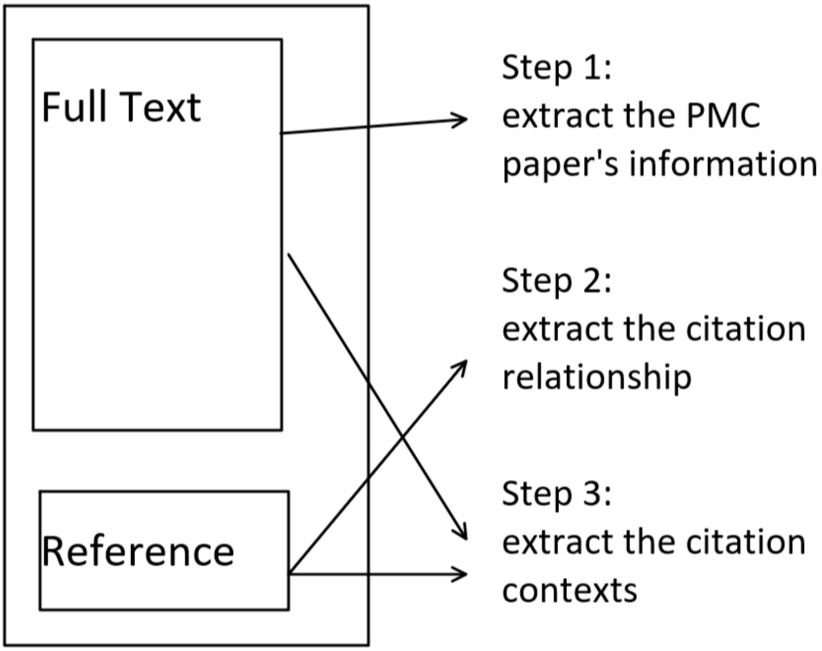


Figure 1. The steps to generate the dataset.

Literature represents a research paper or a scientific document. The PMC paper and its cited documents can all be described using the entity "Literature". To simplify our study, besides the ID of the entity, we chose to save only 4 attributes in our Literature entity. The 4 attributes are: pmc_uid, pmid, local_path (path of the PMC file in the local disk), and fully_updated (a Boolean value, true if a PMC paper is processed). The pmc_uid is used to store the PMC ID or the paper ID for PMC articles, whereas pmid is applied for saving the PubMed ID or the paper ID in the PubMed database. The local_path is the attribute to keep the path of the PMC paper file in the local disk drive, and the attribute fully_updated is a Boolean value used to tag whether a PMC paper is processed or not. Note the fact that only PMC papers have the pmc_uid, local_path and fully_updated, and other documents do not. Also, the documents in the dataset are either PMC citing papers or PMC citing papers' cited documents.

Cite is the entity for storing the relations that one document cites another document. For Cite, there are 4 attributes: citer (the citing document), cited (the cited document), local_reference_id (reference ID for the cited document in the PMC file), and reference_sequence (the sequence for the cited document). Note the fact that both citer and cited are Literature objects, and all the citers are PMC papers in our dataset. But the cited documents include both PubMed documents and non-PubMed documents. The relationship is revealed in Figure 2.

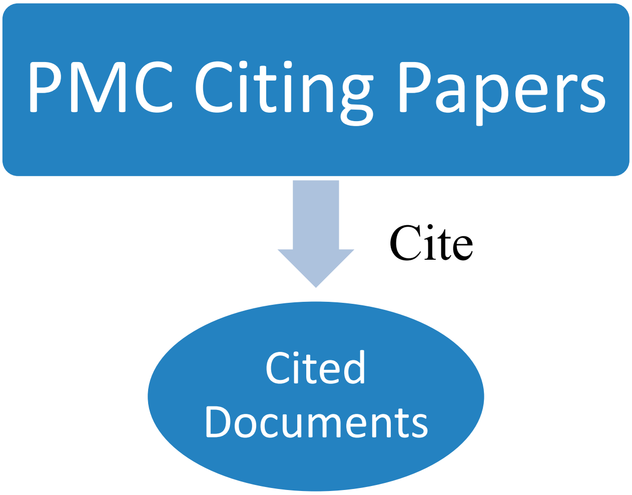


Figure 2. The PMC papers cite cited documents.

CitationContextText is used to save the text of the citation context. As one cited document can be cited multiple times from one citing document and one citation context might have multiple citation markers, we use the CitationContext to link the Cite and CitationContextText. In CitationContext, there are one Cite attribute cite and one CitationContextText attribute citation_context_text. For a citing-cited pair, there can be multiple CitationContext with the same Cite object but different CitationContextText objects. Besides, there is also an attribute called "position" in CitationContext, and we use it to record the exact index of the citation marker in the text of the citation context. All these 4 entities (Literature, Cite, CitationContext and CitationContextText), are represented by class in the object-oriented programming language. Since cite-paragraph is one special case of citation context, we also set two subclasses, CiteParagraph as the subclass for CitationContext and CiteParagraphText as the subclass for CitationContextText.

Some might wonder why we did not use text as an attribute of CitationContext and chose to create a separate CitationContextText to save the text of CitationContext. This is because one paragraph might have multiple CitationContext, and different CitationContext might have the same text of a cite-paragraph. If we use the attribute "text" in CitationContext to save the text, we might have a lot of duplicate data in the database. By using a separate CitationContextText, the same text can be shared by different CitationContext, thus minimizing the size of the database.

PMC paper's metadata

The program first iterates every PMC file in the uncompressed working directory. For each PMC paper, we created a corresponding Literature record in the database. We add the three features extracted from the PMC XML file to the literature object in our database, namely the pmc_uid, pmid and the local_path. The name for each PMC file is in the format of "PMC[PMC ID].nxml" (e.g. PMC100320.nxml). PMC ID (or pmc_uid) is the unique identifier for each PMC paper, and we simply extract the PMC ID from the file name. We search the database if there are any Literature objects with the PMC ID, if there is and fully_updated attribute of the Literature object is also true, then the program will continue to process the next PMC file. If there is none, the PMC file will be processed. Similarly, the local path of the file in our disk can also be extracted for each PMC file in this step.

As the files of PMC papers are in the format of XML (Extensible Markup Language), we can locate the specific attributes with the corresponding XML tag in the PMC file. For a PMC paper, it may have three unique identifiers, namely the PMC ID (or pmc_uid), the PubMed ID (or pmid), and the Digital Object Identifier (DOI). With the help of the XML tag, we can find the pmid for the PMC paper if it is available in the full text. In the dataset, we only focus the pmid and the pmc_uid. We search the database to find out if there is a Literature object with the pmid. If there is, that means it has already been added to the database. This happens when the paper is the cited document for a citing PMC paper that has been processed already. Instead of creating a new object, we update the Literature object with the information in the PMC file; if there is none in the database, we will create a new Literature object and add the information extracted from the PMC file to the newly created object.

Citation relationships

After the elements have been added to the Literature object, our program then extracts the cited documents for the PMC paper, revealing its citation relationships. The cited documents reside in the reference list of the PMC paper. We locate each cited document in reference list and iterate each of the cited documents to obtain the information. Each cited document has a sequence number and a reference ID (such as B12). The sequence number range from 1 to n, where n equals to the number of the cited documents for the paper. As the reference ID not only shows in the reference list, but also appears in the full text where the citation marker is, we can use it to link the cited document and the corresponding citation marker for later citation context extraction. Besides the sequence number and the reference ID, some of the cited documents also have pmid, which also means they are PubMed documents. For these PubMed cited documents, again, we search the database to find if there is a Literature object with the pmid. If there is, we retrieve the Literature object with the pmid as the cited document; if not, we create a new Literature object with the pmid. After the information of the cited document has been gathered, we create a new cite object, where the attribute citer is the PMC paper, and the attribute cited is the cited document. Besides, one Cite object also contains the sequence number and the reference id of the cited document.

Citation context

When the cited documents are added to the database, the next step is to obtain the citation context using the full text of the PMC paper. As we noted before, the reference ID appears both in the cited document in the reference list, and in the full text where the citation marker resides. This feature provides convenient way to help us locate where the citation marker is or where the citation context starts. For each citation marker in the full text, we replace the tag which contains the reference ID (and also the text between the tag) with a new citation marker "~~[sequence number]~~" (e.g. ~~B12~~), where the sequence number equals to the sequence number of the cited document in the reference list. After that, we extract the citation context corresponding to each of the cited document and save the index of the citation marker in the paragraph.

Even though the cited text span is around the citation marker, however, to determine the right window size of the citation context can be a tricky problem (Jha et al., 2017), and the cited text span might vary for different tasks. To simplify the task, some studies use the sentence where the citation marker is in, or the citance (Nakov et al., 2004), as the cited text span (Bornmann, Haunschild, & Hug, 2018; Small, 2018; Small et al., 2017). In our dataset, we assume the window size of the citation context would not exceed the paragraph (or the cite-paragraph) in which the citation marker is in. As a paragraph contains one or more sentences usually dealing with one specific idea, the cite-paragraph is more likely to cover the entire citation context compared with a single sentence. However, in most cases, the cite-paragraph might be much larger than the accurate citation context of one citation marker.

We chose to keep the cite-paragraph as citation context in our dataset, because we want to leave the decisions to be made by future users for further processing. They can choose to implement their technical plan to have a better cited text span for their own use cases. As the citation context does not exceed the cite-paragraph based on our assumption, we view the cite-paragraph as the largest citation context we could possibly have. Besides, we also save the position of the citation marker in the cite-paragraph, which is a number describing how many characters between the starting character of the paragraph and citation marker. We assume the citation markers separated by "and/or" and comma have same position in the cite-paragraph. In the result of the cite-paragraph, to have a clean text for information processing, the citation markers, the "and/or" and commas between the citation markers are all deleted. Therefore, the text in the cite-paragraph might be a little bit different with the original text.

Data records

Our data is saved using MongoDB, which is the database based on distributed file storage. It is designed to provide scalable, high-performance data storage solutions for web applications. Besides, MongoDB is also a database product between the traditional relational database and the non-relational database. It is one of the most feature-rich among non-relational databases and most like a relational database, and suitable for our task. In MongoDB, a table in relational database is called a "collection" and a record in relational database is called a "document". In order not to get confused with the scientific document for our readers, we call each record in MongoDB database as "MongoDB document". In total, there are 37,782,343 Literature objects, 97,526,947 Cite objects, 148,151,328 CitationContext objects, 34,728,502 CitationContextText objects in our database.

Table 1. Literature collection example.

_id ObjectId	*pmc_uid String*	*pmid String*	*local_path String*	*fully_updated Boolean*
5daac89e98fcf8f911522e7a	"2572131"	"18973706"	"/comm_use.A-B.xml/Asia_Pac_Fam_Med/PMC2572131.nxml"	true
5daac89f98fcf8f911522f3e	"2572132"	"18973707"	"/comm_use.A-B.xml/Asia_Pac_Fam_Med/PMC2572132.nxml"	true
5daac89e98fcf8f911522e7b	No field	No field	No field	No field
5daac89e98fcf8f911522e81	No field	"9381230"	No field	No field

In the database, one Literature object have 4 attributes, namely pmc_uid, pmid, local_path and fully_updated. Among the Literature objects in our database, there are 2,658,541 with pmc_uid and local_path, 12,886,421 with pmid, and 24,735,791 without pmc_uid nor pmid. Note the fact that, the Literature objects with pmc_uid and local_path represent our original PMC papers. And as we noted before, only Literature objects for PMC papers have the attributes: pmc_uid, local_path and fully_updated. If one cited document has the PubMed ID, there is a 'pmid' attribute for that document. If there is no pmid for the cited document, all the 4 attributes of the corresponding Literature object will show "No field". In the following example in Table 1, the first two Literature objects represent the citing PMC papers with pmid, the third Literature object represents a non-PMC cited document, and the fourth Literature object represents a cited document with pmid.

A Cite object has 4 attributes, namely citer, cited, local_reference and reference_sequence. The citer stands for the object ID for the citing document and cited represent the object ID for cited document. Besides, the reference sequence and local reference string are also saved. In our example below (see Table 2), there are three Cite objects, and they have the same citing document with the Object ID "5daac89e98fcf8f911522e7a". This citing document cites three different cited documents. Each cited document has different Object ID, reference id ("B1", "B2" and "B3") and sequence ranging from 1 to 3.

Table 2. Cite collection example.

_id ObjectId	*citer ObjectId*	*cited ObjectId*	*local_reference_id String*	*reference_sequence Int32*
5daac89e98fcf8f911522e7c	5daac89e98fcf8f911522e7a	5daac89e98fcf8f911522e7b	"B1"	1
5daac89e98fcf8f911522e7e	5daac89e98fcf8f911522e7a	5daac89e98fcf8f911522e7d	"B2"	2
5daac89e98fcf8f911522e80	5daac89e98fcf8f911522e7a	5daac89e98fcf8f911522e7f	"B3"	3

Each CitationContextText (or CiteParagraphText in our case) has the attribute "text" for its corresponding CitationContext (or CiteParagraph). The text is for the paragraph where the citation context belongs to. Note the fact CiteParagraphText is one special case (or subclass) for CitationContextText. MongoDB use the field _cls to identify which subclass the object belongs to. In the example below (see Table 3), there are two CitationContextText objects.

Table 3. CitationContextText collection example.

_id ObjectId	*_cls String*	*text String*
5e0dc026f40c0c510c201d03	CitationContextText.CiteParagraphText	The placement of a partially … shorter stent patency.
5e0dc026f40c0c510c201d18	CitationContextText.CiteParagraphText	Two types of SEMS were … been previously reported.

As CitationContext is used to link a Cite object and a CitationContextText object, it has one cite field and one citation_context_text field in it representing a Cite object and a CitationContextText object. Besides, the object also has a field called "position" to save the index of the citation marker of the Cite object in the text string of the corresponding CitationContextText. Since CiteParagraph is the subclass of CitationContext, so each object for CitationContext also has the _cls attribute which equals to "CitationContext. CiteParagraph". In the example below (see Table 4), the three CitationContext objects share the same CitationContextText object but have different Cite objects.

Table 4. CitationContext collection example.

_id ObjectId	_cls String	position Int32	cite ObjectId	citation_context_text ObjectId
5e0dc026f40c0c510c201d04	CitationContext.CiteParagraph	204	5e0dc026f40c0c510c201cd4	5e0dc026f40c0c510c201d03
5e0dc026f40c0c510c201d05	CitationContext.CiteParagraph	204	5e0dc026f40c0c510c201cd6	5e0dc026f40c0c510c201d03
5e0dc026f40c0c510c201d06	CitationContext.CiteParagraph	204	5e0dc026f40c0c510c201cd8	5e0dc026f40c0c510c201d03

After the citation context database for PMC articles is built, we use mongodump command in MongoDB to export our data. This command can export all data in the database to a specified directory. The export of the data is in 8 separate files, while 4 of the files are the data files of the 4 collections in our database (namely citation_context_text, citation_context, cite, literature) with the file extension "bson", 4 of the files are the corresponding metadata file of the collections with the file extension "metadata.json". And these 8 files are the data files of the dataset we present in this paper.

Technical validation

To evaluate the validity of our dataset, we chose to select the most cited document as the sample and manually checked the citation contexts or the text of the cite-paragraph pointing to the most cited document to determine if the records were correct. We sorted the citing-cited pairs data in collection Cite, and we found that the most cited document is the paper with PubMed ID 11846609, cited by 22,833 PMC papers (which is approximately 0.86% the PMC papers in our dataset). The paper's title is "Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta CT}$ Method" (Livak & Schmittgen, 2001), authored by Kenneth J. Livaka and Thomas D. Schmittgen. The paper presents the derivation, assumptions, and applications of the $2^{-\Delta\Delta CT}$ method, and the derivation and applications of two variations of the $2^{-\Delta\Delta CT}$ Method that may be useful in the analysis of real-time, quantitative PCR data.

For the most cited document from PMC papers, we extracted the related data from 3 collections in our MongoDB database, which are Cite, CitationContextText and CitationContext. In total, we found no errors in this sample dataset, however, there are some outliers in the dataset that we need to clarify.

For Cite data, we found the most cited document was cited twice by two different PMC citing papers with different reference number. In general, one document only appears once in the reference list with unique reference number or reference ID. However, the two PMC papers wrongfully give the cited document different reference number in the original full text, therefore our Cite data showed that there were two pairs of Cite objects with the same cited and citing documents but different object ID.

For CitationContext, we found three pairs of replicate data with the same position, Cite ID, and CitationContextText ID. After the scrutiny of the full text, we found out these replicate data happened because the same citation marker appears at the same position according to our criterion. For example, in the original full text of the paper, one citation marker might be around the author's name in the full text, and the citation marker with same ID might be around the published year right next to the first citation marker. However, based on our methods, we replaced the XML tag which contains the reference ID (and the text between the tag) with the new unified citation marker, and we also deleted the content between the citation marker in the cite-paragraph. Thus, the citation marker with same reference ID next to each other in the original full text are treated to have same position. The users should be cautious about the issue when using the dataset.

Usage Notes

The data can be accessed freely with the DOI: www.dx.doi.org/10.11922/sciencedb.00393. The users should apply the command mongorestore to load data from the downloaded binary database directory to MongoDB database (see https://docs.mongodb.com/database-tools/mongorestore/). The database directory contains the following files:

- citation_context.bson
- citation_context.metadata.json
- citation_context_text.bson
- citation_context_text.metadata.json
- cite.bson
- cite.metadata.json
- literature.bson
- literature.metadata.json

After the data is loaded in MongoDB by the command mongorestore, the users can then use tools provided by MongoDB to analyze the data.

Limitations

The citation context dataset based on PubMed Central papers we provided is one of the largest citation context datasets, however, the dataset has several limitations.

Firstly, to simplify the processing of the original data, only the relationships between the PubMed Central papers and PubMed documents are revealed. For a document without PubMed ID in the database, there is only one citing PMC paper for it, but in fact, there may be more citing PMC papers for the document. The next version of the dataset might utilize the DOI in the full text and other third-party citation services to cover cited documents without PubMed ID.

Secondly, for a Literature object, there are only four attributes, namely pmc_uid, pmid, local_path, fully_updated, and only pmc_uid and pmid are the metadata of the paper. In the future, we intend to use Entrez Programming Utilities at the National Center for Biotechnology Information to add more metadata (such as the title, authors, publication details) for Literature object. With more metadata of a document, more interesting insights might be discovered from the dataset.

Lastly, unlike general citation database such as the Web of Science, Scopus, Microsoft Academic, Google Scholar, the original PubMed Central Open Access Subset only covers a small part of the papers published, and the papers in the subset are in the field of biomedical and life sciences. The users of the dataset for citation analysis should be cautious about the scope of the dataset and ignorance of the issue might lead to biased conclusion.

Acknowledgments

This paper is funded by the following:

1. Liaoning Planning Office of Philosophy and Social Science (No. L20BGL003)
2. Department of Education of Liaoning Province (Liaoning Education Office [2020] No. 122)
3. Natural Science Foundation of Liaoning Province (No. 2020-MS-317)
4. Liaoning Planning Office of Philosophy and Social Science (No. L19BXW010)
5. Doctoral Research Start-up Fund of Liaoning Province (No. 2019-BS-014)
6. Liaoning Province 2018 Key R&D Guidance Program Projects (No. DNUI01)

References

Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M.-Y., Lee, D., et al. (2008). The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. EUROPEAN LANGUAGE RESOURCES ASSOC-ELRA.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80.

Bornmann, L., Haunschild, R., & Hug, S. E. (2018). Visualizing the context of citations referencing papers published by Eugene Garfield: A new type of keyword co-occurrence analysis. *Scientometrics*, *114*(2), 427–437. Springer Netherlands.

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*(1), 59–73. Elsevier Ltd.

Bradshaw, S. (2002). Reference directed indexing: Indexing scientific literature in the context of its use. Northwestern University.

Bradshaw, S. (2003). Reference directed indexing: Redeeming relevance for subject search in citation indexes. *International conference on theory and practice of digital libraries* (pp. 499–510). Berlin, Heidelberg.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, *65*(9), 1820–1833.

Färber, M., Thiemann, A., & Jatowt, A. (2018). A high-quality gold standard for citation-based tasks. Presented at the Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Gipp, B., Meuschke, N., & Lipinski, M. (2015). CITREC : An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central.

He, J., & Chen, C. (2018). Temporal Representations of Citations for Understanding the Changing Roles of Scientific Publications. *Frontiers in Research Metrics and Analytics*, *3*(September), 1–9.

Huang, W., Wu, Z., Liang, C., Mitra, P., & Giles, C. L. (2015). A Neural Probabilistic Model for Context Based Citation Recommendation. *Twenty-ninth Aaai Conference on Artificial Intelligence*.

Jha, R., Jbara, A. A., Qazvinian, V., & Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, *23*(1), 93–130.

Lawrence, S., Giles, C. L., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *Computer*, *32*(6), 67–71.

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. *Methods*, *25*(4), 402–408.

Nakov, P. I., Schwartz, A. S., & Hearst, M. A. (2004). Citances: Citation Sentences for Semantic Analysis of Bioscience Text. *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics* (pp. 1–8).

O'Connor, J. (1982). Citing statements: Computer recognition and use to improve retrieval. *Information Processing and Management*, *18*(3), 125–131. Elsevier.

O'Connor, J. (1983). Biomedical citing statements: Computer recognition and use to aid full-text retrieval. *Information Processing and Management*, *19*(6), 361–368.

Radev, D. R., Muthukrishnan, P., Qazvinian, V., & Abu-Jbara, A. (2013). The ACL anthology network corpus. *Language Resources and Evaluation*, *47*(4), 919–944. Springer.

Ritchie, A. (2008). *Citation context analysis for information retrieval* (PhD Thesis). *PhD thesis*. University of Cambridge.

Roberts, R. J. (2002). PubMed Central: The GenBank of the published literature. *Proceedings of the National Academy of Sciences*.

Saier, T., & Färber, M. (2020). unarXive: A large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*, *125*.

Small, H. (1978). *Cited documents as doncept symbols*. *Social Studies of Science* (Vol. 8).

Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *Journal of Informetrics*, *12*(2), 461–480. Elsevier Ltd.

Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics*, *11*(1), 46–62. Elsevier Ltd.

Sugiyama, K., & Kan, M. Y. (2015). A comprehensive evaluation of scholarly paper recommendation using potential citation papers. *International Journal on Digital Libraries*, *16*(2), 91–109.

Sun, F. (2020, December 18). PubMed Central Citation Context Dataset V1. Science Data Bank. Retrieved from www.dx.doi.org/10.11922/sciencedb.00393

Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, *64*(7), 1490–1503.

Visualizing Characteristic Interdisciplinary Research Collaborations Among Departments in a University

Tetsuya Takahashi[1], Marie Katsurai[1], Ikki Ohmukai[2] and Hideaki Takeda[3]

[1] *{takahashi, katsurai}@mm.doshisha.ac.jp*
Doshisha University, 1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto 610-0394 (Japan)

[2] *i2k@l.u-tokyo.ac.jp*
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033 (Japan)

[3] *takeda@nii.ac.jp*
National Institute of Infomatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 (Japan)

Abstract

This research-in-progress paper presents a novel visualization approach to facilitate understanding of interdisciplinary collaboration within a university using a large-scale dataset of research grant projects in Japanese universities. First, we construct a network of departments in a target university, and then visualize the activeness of inter-departmental collaborations using member information of research projects. Second, to compare research features among universities, we quantify the difference in the frequency of research field combinations between the target university and other universities. Collaborations that have characteristic research field combinations are then highlighted in the network, which can be useful for the target university's research promotion strategies. We present a case study at the University of Tokyo to validate the effectiveness of our approach. To enhance our visualization's advantage, we also demonstrate that the average amount of research grants that the characteristic research field combinations received is significantly higher than those of field combinations that are prevalent among several universities.

Introduction

Interdisciplinary collaborations have become increasingly important for providing innovative solutions to complex problems. Research administrators at a university often make efforts to analyze the performance of its research institutes to create an interdisciplinary team from different departments. Understanding the characteristics of research at the university compared with other universities is also crucial for choosing research promotion strategies. However, there is no established visualization framework for such a comparison of research features among universities.

Previous studies have defined measures of interdisciplinarity for several academic entities, such as journals and authors. For example, Rodríguez (2017) proposed a citation-based indicator for classifying scientific journals into four classes based on their degree of interdisciplinarity. Zhang et al., (2020) used a co-author network to assess the interdisciplinarity of each researcher. To examine the status of collaborations across research topics, Rafols and Meyer (2010) evaluated the degree of interdisciplinarity of papers published in bioscience and visualized them on a two-dimensional plane. Some studies focused on using the attributes of researchers listed in individual research projects. L. Zhang et al., (2018) analyzed the diversity of disciplines estimated from the authors' varied affiliations. Abramo, D'Angelo, and Di Costa (2012) analyzed collaboration frequencies among researchers from different fields to identify frequent field combinations. Similarly, Uddin, Imam, and Mozumdar (2020) constructed a network in which each node represents a discipline, while an edge between two nodes represents the participation of corresponding disciplines in grant projects. The line of these science mapping research projects has the potential to be extended to institutional research in a university.

In this research-in-progress paper, we present a novel visualization approach to facilitate understanding of a university's collaboration characteristics in interdisciplinary research. Our study uses a large-scale dataset of research grant projects in Japan. First, we construct a network of departments using member information of research projects to show the activeness of inter-departmental collaborations. Second, to compare universities, we quantify the difference in frequency of research field combinations between the target university and other universities. Collaborations that have characteristic research field combinations are then highlighted in the network, which can be useful for the target university's research promotion strategies. A case study at the University of Tokyo is presented in this paper to validate the effectiveness of our approach. We also investigate the difference in the average amount of research grants received by field combinations judged as characteristic and those deemed common.

Dataset

Our study requires a dataset containing collaborative research project information, including member affiliations and research field labels. To construct such a dataset, we chose KAKEN[1] as an information source. KAKEN is the database of Grants-in-Aid for Scientific Research projects granted by the Ministry of Education, Culture, Sports, Science and Technology and the Japan Society for the Promotion of Science for research projects in all fields in Japan. Among KAKEN research projects from fiscal year (FY) 2004 to FY 2017, we extracted the projects that have at least two researchers and classified them as "collaborative research projects." In KAKEN, each research project is classified into a single field category by the Principal Investigator according to the application year's list of categories. We first manually integrated these field categories, which differ from year to year, into the 13 field labels of the list as of 2017. Although field labels are provided to research projects in KAKEN, and not to individual researchers, our study requires collaborators' field information similar to that used in Abramo, D'Angelo, and Di Costa's study (2012). Thus, we assigned the field label of a research project to the project's Principal Investigator so that each researcher can be associated with one or more field labels. The resulting dataset consisted of 112,722 research projects, 4,054 domestic institutions, and 130,733 unique researchers.

Measures of a university's collaborative research

Activeness of collaborative research

We first aim to investigate the activeness of collaborations between different departments in a target university. Let U be the set of all universities in the dataset. The set of departments in a target university $u \in U$ is denoted by Ω_u. Let $T^u_{d,d'}$ be the set of research projects, which contain at least one researcher from department $d \in \Omega_u$ as well as at least one researcher from department $d' \in \Omega_u$. We represent how active the inter-departmental collaborations between departments d and d' using the cardinality of the project set as follows:

$$Active_{collab}(d,d') = \left|T^u_{d,d'}\right|. \quad (1)$$

Although this simple count-based measurement contributes to finding frequent patterns of inter-departmental collaborations "within" a target university, it cannot reveal how characteristic each pattern is, "compared with" other universities. We solve this problem in the following subsection.

[1] https://kaken.nii.ac.jp/en/

Activeness of characteristic interdisciplinary collaborative research

To characterize features of the university's inter-departmental collaborations, we define a combinatorial set of field categories of researchers participating in a single research project as a "*field combination*." For example, suppose a project is conducted by four researchers whose fields are computer science (CS), humanities (H), CS, and engineering (E), respectively, we can extract from this project the following field combination; (CS, H, E). Note that if a researcher has multiple labels, all the labels are considered to create a combination. We assume that if a certain field combination occurs frequently in a target university compared with other universities, the corresponding topic can be the characteristic of the university. Specifically, the characteristic of a field combination c for a target university $u \in U$ is quantified using the following measure:

$$Measure_{character}(u,c) = t\_ratio(u,c) - \frac{1}{|U-\{u\}|}\sum\nolimits_{j \in U-\{u\}} t\_ratio(j,c)\,, \qquad (2)$$

$$t\_ratio(j,c) = \frac{\text{Number of research projects corresponding to field combination } c \text{ at university } j}{\text{Total number of research projects at university } j}\,.$$

A large value of $Measure_{character}(u,c)$ means the field combination c can represent characteristic collaboration at university u.

Next, we quantify the activeness of "characteristic" interdisciplinary collaborations between departments in a target university $u \in U$. The field combination extracted from project $t \in T^{u}_{d,d'}$ is denoted by $c(t)$. For two departments d, d', focusing on how many of their collaborations were produced by characteristic field combinations, we compute the following activeness measure:

$$Active_{character}(d,d') = \sum\nolimits_{t \in T^{u}_{d,d'}} g(c(t)) \;\; (3),$$

$$g(c) = \begin{cases} Measure_{character}(u,c), & \text{if } Measure_{character}(u,c) > \mu_{u,c} + \sigma_{u,c}, \\ 0, & \text{if } Measure_{character}(u,c) \le \mu_{u,c} + \sigma_{u,c}, \end{cases}$$

where $\mu_{u,c}$ and $\sigma_{u,c}$ are the mean and the standard deviation of the values that satisfy $Measure_{character}(u,c) > 0$, respectively. If $Measure_{character}(u,c(t)) > \mu_{u,c(t)} + \sigma_{u,c(t)}$, the project t is defined as a "characteristic" interdisciplinary collaboration, and its activeness score contributes to $Active_{character}(d,d')$. A large value of $Active_{character}(d,d')$ means that the two departments d and d'actively collaborate, especially in research topics of characteristic field combinations for the university. Using this measurement, we can easily find which department pair conducts collaborations that can be features of the university.

Results

Visualization based on network construction

To verify the effectiveness of the proposed method, we present a case study at the University of Tokyo, which has 13 research departments. We first constructed two networks whose nodes and edges are departments and their collaboration relationships, respectively. In each network, nodes and edges are weighted using Eq. (1) or Eq. (3). Note that for $d = d'$, each measure represents the activeness of "intra-departmental" collaborations in department d, which was used as node weights. Figure 1 shows the network weighted using Eq. (1), in which each edge/node is colored dark when the corresponding weight is high, with a light color used when the value is low. The figure shows that most of collaborations occur within departments (i.e., $d = d'$) and that many collaborative research projects are conducted within the Graduate School of Medicine compared to other departments. Figure 2 shows the network weighted using Eq. (3), in which the edge/node is colored, similar to Fig. 1. From Fig. 2, we can see the following two trends: (a) the Graduate School of Engineering actively conducts characteristic interdisciplinary collaboration within the department; and, (b) the Graduate School of Engineering and the Graduate School of Frontier Sciences actively collaborate in research projects belonging to characteristic field combinations.

Although we found from Fig. 1 that a number of collaborative research projects are conducted within the Graduate School of Medicine, Fig. 2 implies that their research field combinations are similar to other universities. For the Graduate School of Science, Fig. 1 shows that the number of collaborative research projects within the department is not large, while Fig. 2 indicates that researchers in the Graduate School of Science conduct many “characteristic” interdisciplinary collaborations within the department. We can consider that this department has the potential to create interdisciplinary collaboration teams with other the departments.

In addition, Table 1 shows the field combinations evaluated as characteristic interdisciplinary collaborations, sorted in descending order of $Measure_{character}$. We can see that environmental science/engineering is the most characteristic field combination in the University of Tokyo. The table shows that engineering is included in the higher rank of field combinations. We can confirm that the Graduate School of Engineering, to which many researchers assigned under the Engineering label are considered to belong, indeed has darker coloration of nodes and edges from Fig. 2. These findings will facilitate mapping a strategy to promote characteristic collaborations in the university.

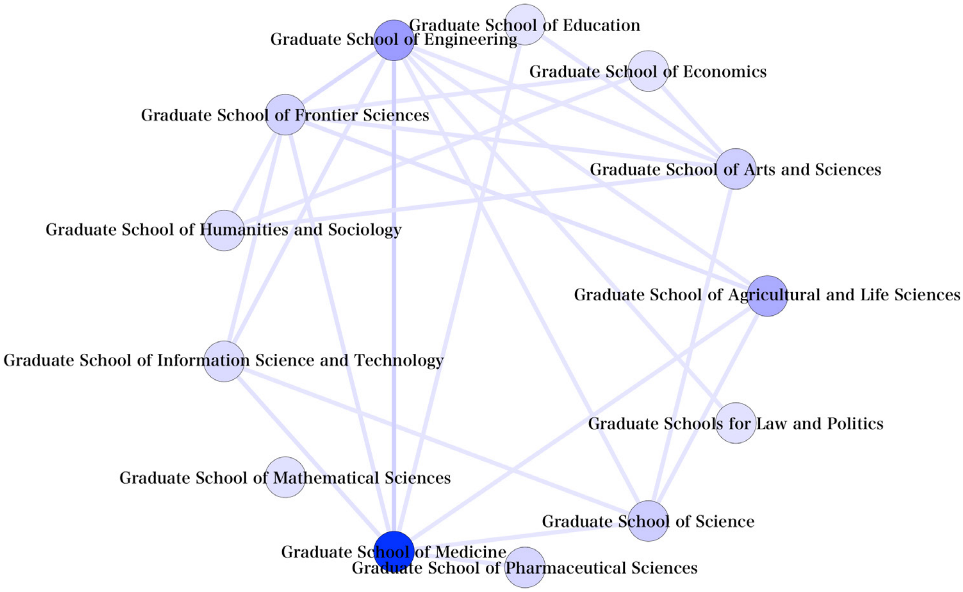

Graduate School of Education
Graduate School of Engineering
Graduate School of Economics
Graduate School of Frontier Sciences
Graduate School of Arts and Sciences
Graduate School of Humanities and Sociology
Graduate School of Agricultural and Life Sciences
Graduate School of Information Science and Technology
Graduate Schools for Law and Politics
Graduate School of Mathematical Sciences
Graduate School of Science
Graduate School of Medicine
Graduate School of Pharmaceutical Sciences

Figure 1. Visualization of the activeness of research collaborations based on Eq. (1).

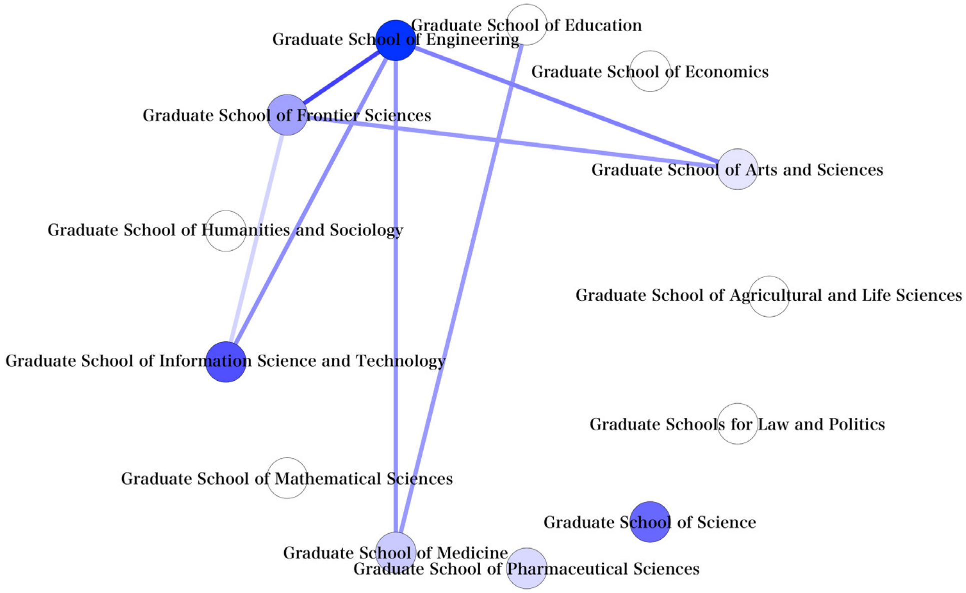

Graduate School of Education
Graduate School of Engineering
Graduate School of Economics
Graduate School of Frontier Sciences
Graduate School of Arts and Sciences
Graduate School of Humanities and Sociology
Graduate School of Agricultural and Life Sciences
Graduate School of Information Science and Technology
Graduate Schools for Law and Politics
Graduate School of Mathematical Sciences
Graduate School of Science
Graduate School of Medicine
Graduate School of Pharmaceutical Sciences

Figure 2. Visualization of the activeness of "characteristic" collaborations based on Eq. (3).

Table 1. Characteristic field combination in the University of Tokyo.

Rank	*Field combination*
1	Environmental science / Engineering
2	Interdisciplinary science and engineering / Engineering
3	Complex systems / Medicine, dentistry, and pharmacy / Chemistry
4	Interdisciplinary science and engineering / Mathematical and physical sciences
5	Complex systems / Medicine, dentistry, and pharmacy / Biological Sciences
6	Interdisciplinary science and engineering / Mathematical and physical sciences / Engineering
7	Interdisciplinary science and engineering / Chemistry / Engineering
8	Engineering / Informatics
9	Biology / Medicine, dentistry, and pharmacy / Biological Sciences
10	Environmental science / Mathematical and physical sciences

The impact of characteristic interdisciplinary collaborations on grant budgets

We further conducted in-depth analyses of the "characteristic" interdisciplinary collaborations found by our approach. Research projects in KAKEN were provided with budgets, whose amounts differ from one another. In this experiment, we compare the budgets' amounts between the research projects that are considered to be characteristic interdisciplinary collaborations and those that are not. Our research question here is: *Can characteristic interdisciplinary collaborations obtain more research budgets than non-characteristic ones?*

Given that the range of budget allocations for research projects varies, depending on the grant application categories, this experiment used only research projects in the category of *Grant-in-Aid for Scientific Research (B)*, which generally presupposes collaboration with multiple researchers. Projects in Grant-in-Aid for Scientific Research (B) are supported by budgets ranging between 5 to 20 million yen. The average of the allocations for research projects belonging to different types of collaboration is shown in Table 2. We used the Mann-Whitney U test to examine whether there is a difference between the two groups. The two-sided test showed that the significance probability p was $p = 0.019 < 0.05$. Interestingly, it is demonstrated that the budget average amount allocated to the research projects judged as characteristic is significantly higher than that allocated to the research projects considered common.

Table 2. Comparison of average amounts of budget allocated to projects.

	Judged as	
	characteristic	*common*
Number of research project*s*	45	131
The average of the allocations (yen)	16,650,666	15,786,564

Conclusion

This research-in-progress paper proposed a novel visualization that facilitates better understanding of interdisciplinary collaborations within a university. We constructed a network representing the activeness of inter-departmental collaboration within a target university. We then found the characteristic field combination in the target university, and the activeness of the characteristic interdisciplinary collaboration is represented in the same network form. In addition, we demonstrated that the average of research grants that the characteristic research field combinations received is significantly higher than that obtained by common ones. In future work, we will apply the method to other research institutions for large-scale comparisons. We will also develop an interdepartmental collaborator recommendation system based on the proposed method like in our previous study (Takahashi, Tango, Chikazawa, & Katsurai, 2020).

Acknowledgments

This research was partly supported by JSPS KAKENHI Grant Number 20H04484, JST ACT-X grant number JPMJAX1909, and ROIS NII Open Collaborative Research 2020-20S0405.

References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2012). Identifying interdisciplinarity through the disciplinary classification of coauthors of scientific publications. *Journal of the American Society for Information Science and Technology*, *63*(11), 2206–2222.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, *82*(2), 263–287.

Rodríguez, J. M. (2017). Disciplinarity and interdisciplinarity in citation and reference dimensions: knowledge importation and exportation taxonomy of journals. *Scientometrics*, *110*(2), 617–642.

Takahashi, T., Tango, K., Chikazawa, Y., & Katsurai, M. (2020). A Novel Researcher Search System Based on Research Content Similarity and Geographic Information. *ICADL 2020: Digital Libraries at Times of Massive Social Transition*, *12504 LNCS*, 390–398.

Uddin, S., Imam, T., & Mozumdar, M. (2020). Research interdisciplinarity: STEM versus non-STEM. *Scientometrics*, 1–16.

Zhang, L., Sun, B., Chinchilla-Rodríguez, Z., Chen, L., & Huang, Y. (2018). Interdisciplinarity and collaboration: on the relationship between disciplinary diversity in departmental affiliations and reference lists. *Scientometrics*, *117*(1), 271–291.

Zhang, W., Shi, S., Huang, X., Zhang, S., Yao, P., & Qiu, Y. (2020). The distinctiveness of author interdisciplinarity: A long-neglected issue in research on interdisciplinarity. *Journal of Information Science*.

Editorial board member and prolific author status positively shorten publication time

Zehra Taşkın[1], Abdülkadir Taşkın[2], Güleda Doğan[3] and Emanuel Kulczycki[1]

[1] *{zehra.taskin, emek}@amu.edu.pl*
Scholarly Communication Research Group, Adam Mickiewicz University in Poznań (Poland)

2 *abdulkadir.taskin@gmail.com*
Independent researcher (Poland)

3 *gduzyol@hacettepe.edu.tr*
Department of Information Management, Hacettepe University (Turkey)

Abstract

Publication speed is one of the important aspects of scholarly communication today since a good deal of research performance evaluation systems are based on published articles. This study aims to reveal the factors affecting the publication speed of journals. In this context, six library and information science (LIS) journals, *ASLIB Journal of Information Management, Journal of Documentation, Journal of Informetrics, Journal of the Association for Information Science and Technology, Online Information Review,* and *Scientometrics* are analysed in terms of their publication speed. Results show that being an editorial board member or prolific author for journals significantly shortens the duration of publication. Moreover, when there is at least one editorial board member or prolific author in the author group, the duration of the publication is shorter than the articles from the unknown authors. However, the fact that no significant difference is determined between single- and double-blind peer review and the duration of publication process gives an idea about the scientific levels of articles written by editorial board members or prolific authors. In this regard, our approach is to examine other factors affecting the publication speed by conducting multi-dimensional analysis in future studies.

Introduction

Peer review, which can be carried out blind (single or double) or open (the identities of reviewers are disclosed at least to the authors), is one of the most effective tools used to make decisions on research quality. However, it has various drawbacks. Most of the problems of current peer review systems are related to the reliability of reviewers' ratings and biases such as status bias (institutional or individual level) or gender bias (Cox et al., 1993, p. 313).

Comparisons for measuring the effectiveness of single- and double-blind review systems are important to choose a fairer and balanced system as well as reduce biases. Blank (1991) showed that the reviewers were more critical of double-blind peer review and the authors from peripheral countries received less acceptance in the double-blind review system. It was revealed in a different study that single-blind reviewers were significantly more likely than their double-blind counterparts to accept papers from popular authors, top universities, or top countries (Tomkins et al., 2017). Okike et al. (2016, p. 1316) concluded that if blinding was not applied properly, variables such as gender or popularity affect acceptance or rejection decisions, and Budden et al. (2008) stated that double-blinding increased the representation of women in science. Sun et al. (2021) suggested a double-blind review system to remove prestige bias in the review process. While Snodgrass (2006, p. 10) indicated that the main problem of the single-blind review was the unfairness to unknown authors, on the other hand, according to some others single-blind reviews provided speed in the peer review process by providing personal knowledge about the authors or works (McCormack, 2009). However, according to a study based on the articles published in *American Economic Review* (Blank 1991, p. 1048), the duration of the single-blind review was two weeks longer than the double-blind.

Studies in the literature have confirmed that there are various practices in all types of peer review processes regarding the journal, gender, country, prestige, etc. The common point of almost all these studies is to reveal who the peer review provides an advantage. In this case, it is possible to say that the concept defined as the Matthew Effect in science, which represents the cumulative advantage of science elites (Merton, 1968), is also valid for peer review processes. According to Merton, reputed or distinguished researchers have more credit than researchers who are unknown in the same field, even if their works were quite similar (p. 59).

It is possible to mention Matthew Effect not only in article acceptance decisions but also in the duration of peer review. While some researchers' articles are published in a very short time, this process may take longer for some other researchers. However, publication speed has importance for all researchers who seek tenure and incentives. The fact that long review processes, the lack of standardization in the processes, and differences from discipline to discipline cause researchers to have negative opinions about the journals (Huisman & Smits, 2017).

The main aim of the paper is to reveal factors affecting the duration of peer review processes for library and information science journals. The research questions are:

- How do the types of peer review affect the duration of peer review?
- Does being an editorial board member of a journal shorten the peer review process?
- Does being one of the most prolific authors of the journal shorten the peer review process?
- Does having more than one role in a journal affect the peer review speed?
- What is the difference between the publication speed of researchers who are central in the LIS field and others?

Data and Methods

To achieve the aim of the study, three single-blind journals of library and information science (LIS) field (*Journal of Informetrics* [JOI],[1] *Journal of the Association for Information Science and Technology* [JASIST][2] and *Scientometrics* [SCIM][3]), and three double-blind journals (*ASLIB Journal of Information Management* [ASLIB],[4] *Journal of Documentation* [JDOC][5] *and Online Information Review* [OIR][6]) were chosen. A total of 3,816 articles that were published between 2016 and 2021 were evaluated deeply. Metadata of articles were gathered from Web of Science on January 2, 2021. In the dataset, a total of 2,843 single and 973 double-blind articles were stored. 45% (1,715) of these articles were published in *SCIM*, 20% (753) in *JASIST*, 10% in *JDOC* (381), *JOI* (375), and *OIR* (367), and 5% (225) in *ASLIB*. Therefore, differences between journals should be considered to interpret the results. To achieve this, the average duration for each journal was presented in figures.

To collect the names of editorial board members, the websites of journals were used. All members of the editorial boards were considered. However, considering the important editorial board change for *JOI* in 2019 (Larivière, 2019), two different lists were used for this journal. One for the publications before 2019 and one for publications after 2019. For finding the most prolific authors of each journal, six different searches were done in Web of Science. The last

[1] https://www.elsevier.com/journals/journal-of-informetrics/1751-1577/guide-for-authors
[2] https://asistdl.onlinelibrary.wiley.com/hub/journal/23301643/homepage/forauthors
[3] https://www.springer.com/journal/11192/submission-guidelines
[4] https://www.emeraldgrouppublishing.com/journal/ajim#author-guidelines
[5] https://www.emeraldgrouppublishing.com/journal/jd#author-guidelines
[6] https://www.emeraldgrouppublishing.com/journal/oir#author-guidelines

10 years (2011-2020) were considered. Only articles and reviews were covered. The first ten authors of each journal were defined as "prolific authors". If the tenth and eleventh authors had the same number of publications, two of them were added to the dataset. After finding the most prolific authors in Web of Science, the articles were classified accordingly.

The codes were written using Python to get publications' timelines automatically from journal websites (Taşkın, 2021). To compare the review durations in terms of the journal and peer review types, Kruskal Wallis, Mann Whitney, and Chi-square tests are applied considering the assumptions required to apply parametric testing are not met. Effect sizes are also calculated for the positive test results. Formula 1 shows the effect size calculation for Kruskal Wallis (η_H^2), and Formula 2 for Mann Whitney U (r_G) tests where H is the Kruskal-Wallis test statistic, k is the number of groups, $\bar{R}_A$ and $\bar{R}_B$ are the average ranks for groups, n and N_T is the total number of observations (Cohen, 2013, p. 10-11, 19-20).

$$\eta_H^2 = (H - k + 1)/(n - k) \qquad \text{(Formula 1)}$$
$$r_G = 2(\bar{R}_A - \bar{R}_B)/N_T \qquad \text{(Formula 2)}$$

We used SPSS (version 21) and RCommander for statistical tests and descriptives; RCommander (with KMggplot2 plugin) and Flourish Studio for visualization.

Findings

Effects of single- and double-blind reviews on the duration of peer review

When it was investigated whether the type of peer review affects the review duration, small differences were found between the two groups (see Figure 1). Although the acceptance periods of the specific journals varied, the average duration of the single- and double-blind review types were very similar. Mann-Whitney U test confirmed this similarity (*U*=1358685.500, *Z*=-0.824, *p*=0.410). On the other hand, significant differences were found between the journals' peer review durations regardless of peer review type. According to the results of the Kruskal Wallis test, significant differences were found for the average duration of peer review (*H*=401.315, p<0.001, η_H^2=0.104). When all journals were compared to understand the source of differences, there were no statistically significant differences found between *ASLIB* & *JOI* and *JDOC* & *JOI* at %99.9 confidence level.

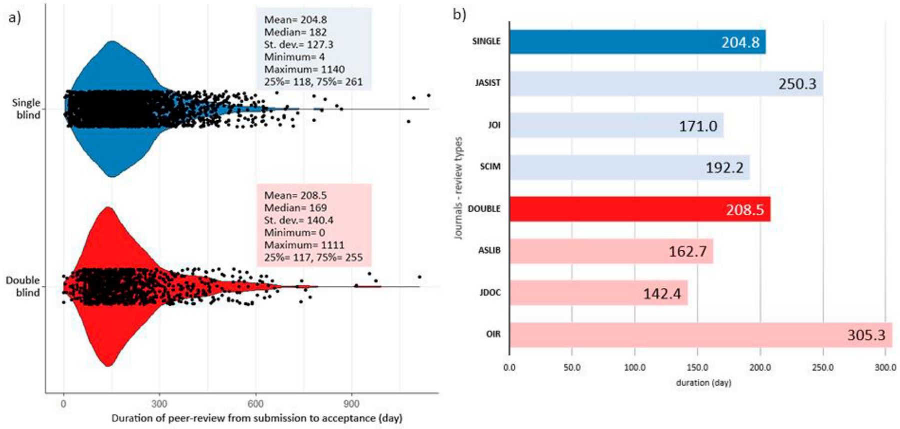
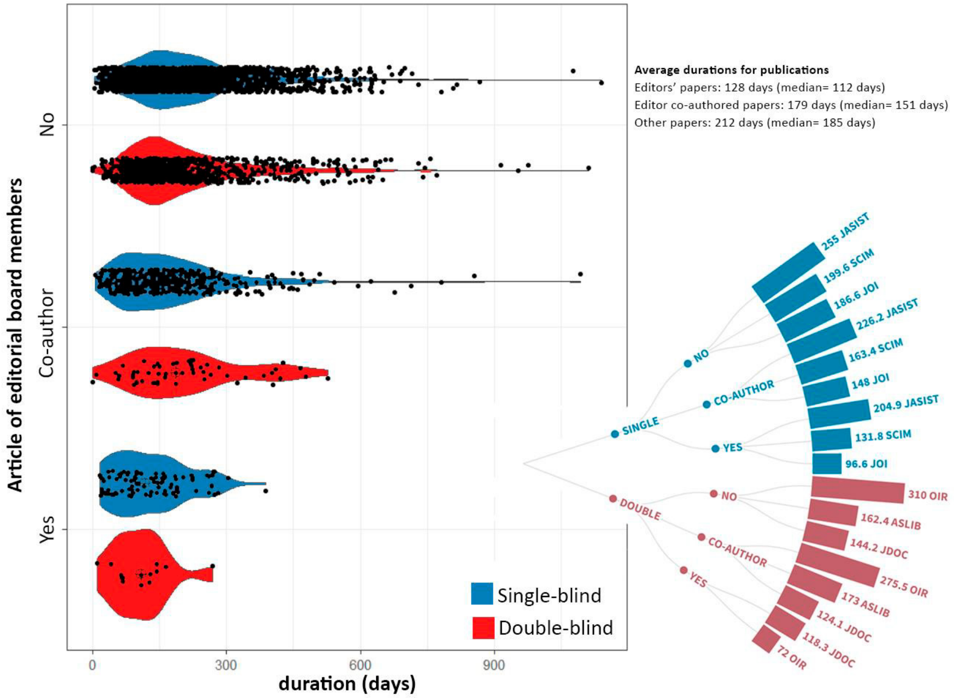


Figure 1. Duration of peer review regarding peer review types

Effects of being an editorial board member on the review durations

Being an editorial board member is important in showing that researchers have achieved a certain scientific level, have proven their scientific merit in the field, and thereby have become a decision-maker for a journal (Bedeian et al., 2009; Pardeck & Meinert, 1999). Members of the editorial boards are often selected among the most popular researchers in their fields with high scientific competencies. Therefore, it is expected that the peer review processes of the articles written by editorial board members may be completed faster in parallel with the recognition, experience, and popularity in the field. The differences between the peer review durations of editorial board members' and other authors' papers (see Figure 2) are statistically significant (H=92.892, p<0.001, η^2_H=0.024).[7] One of the important findings is that, when an editorial board member is a co-author of a study, it makes the peer review process shorter. It confirms a recent study (Zhang et al., 2021) that reveals the positive effects of editorial boards' cooperation with the authors' publications. Furthermore, this difference is more obvious for single-blind journals (η^2_H=0.032).



Average durations for publications
Editors' papers: 128 days (median= 112 days)
Editor co-authored papers: 179 days (median= 151 days)
Other papers: 212 days (median= 185 days)

Figure 2. Peer review duration for papers written by editorial board members

Papers from the prolific authors of the journals

According to Merton (1968, p. 61), the works of scientists who have an outstanding position in science have been validated by judgments of the average quality of their past work. Therefore, it is easier to accept the works of outstanding authors by the journals. To confirm whether this approach affects the review durations, the review durations of the works from the prolific authors of each journal were evaluated in this study (see Fig. 3).

[7] The difference is significant for all the pairs. According the effect sizes calculated, the most significant difference is between the papers of editors (Yes in Figure 2) and papers with no editors (No in Figure 2) (U=106418.000, Z=-7.636, p<0.001, r_G=0.420)

The results show that there are significant differences in peer review durations for the papers from productive authors and others ($H(2)=151.477$, $p<0.001$, $\eta_H^2=0.039$). The effects sizes calculated for each of the three pairs showed that the most significant difference is between the papers of prolific authors (Yes in Figure 3) and papers with no prolific authors (No in Figure 3) ($U=75011.500$, $Z=-9.265$, $p<0.001$, $r_G=0.548$). On the other hand, when the effect of prolific authors on peer review duration evaluated separately for single- and double-blind journals, single-blind journals stand out with the more pronounced difference ($\eta_H^2=0.046$) in comparison with double-blind ones ($\eta_H^2=0.024$).

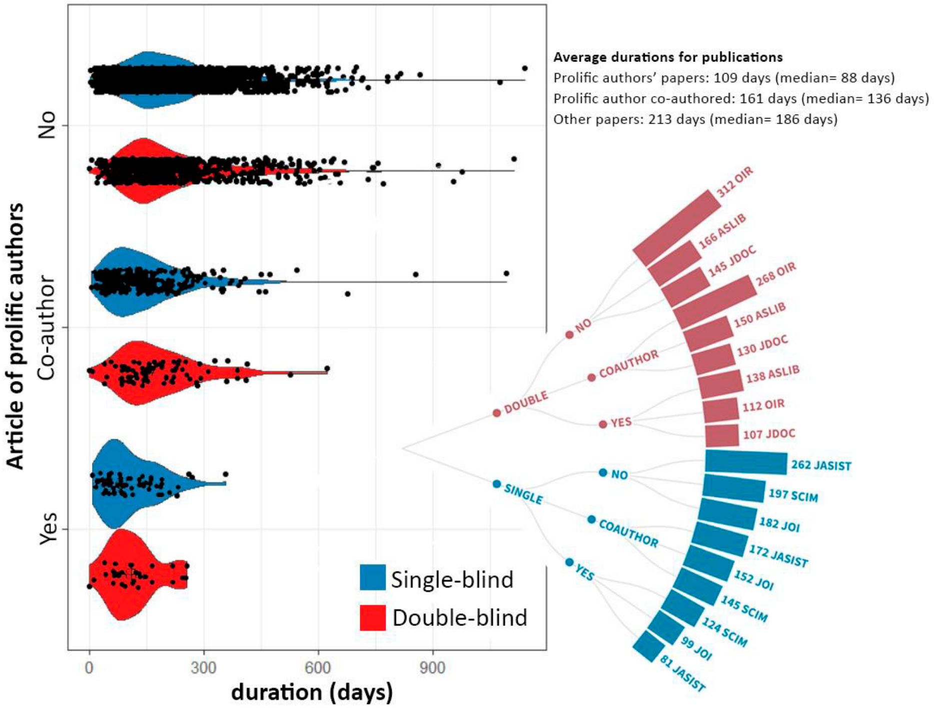


Figure 3. Peer review duration for papers written by prolific authors of the journals in the last 10 years

What if an editorial board member is also a prolific author?

It was revealed that being an editorial board member or a prolific author of a journal shortens the review process. The main question after this finding is that what if an editorial board member is also a prolific author? According to the results (See Figure 4), if the editorial board members in both single and double-blind peer review were also the most prolific authors and submitted their articles to their journals, it took an average of 100 days to complete all processes. Having one editorial board member or a prolific author as a co-author may also shorten the process. The longest evaluation process was identified for the papers written by unknown researchers of the journals.

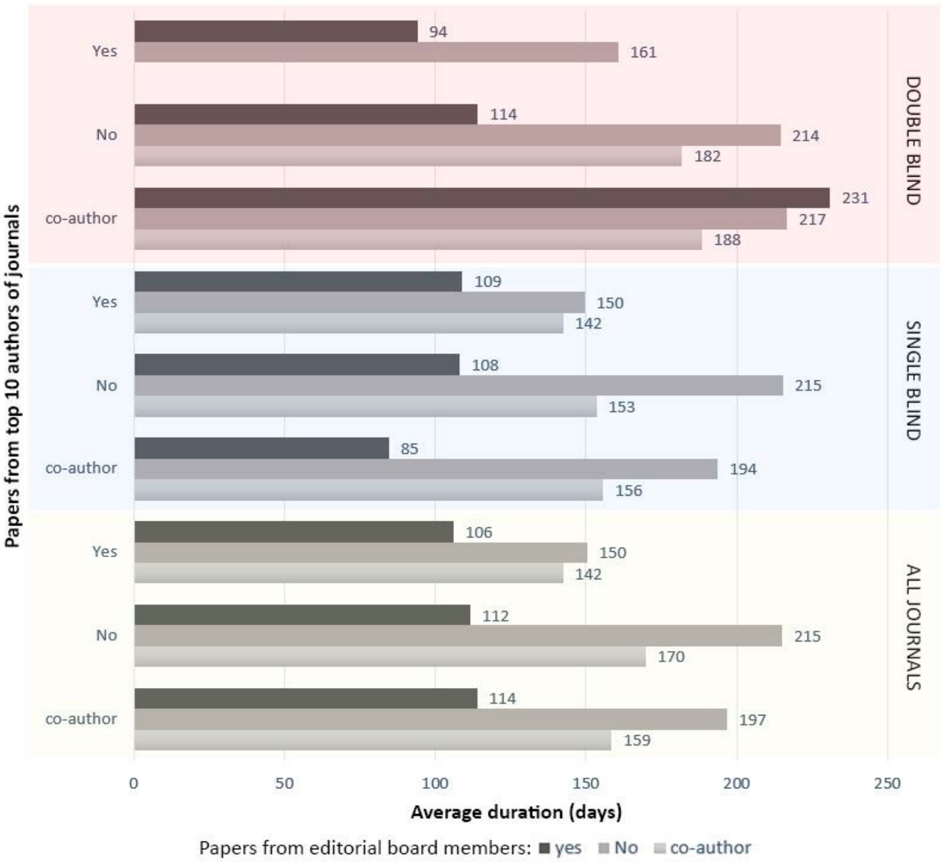


Figure 4. Average durations for the papers written by editorial board members and prolific authors

Peer review durations of central researchers of the LIS field

All the findings of previous parts have proved that there are popular, successful, and experienced researchers working in the LIS field, and their papers published faster than the others, as expected. Looking from the wider perspective, a total of 6,814 unique authors published 3,816 papers in LIS. Some authors serve more than one journal as an editorial board member. Besides, most of the prolific authors of single-blind journals are also editorial board members. Figure 5 shows the distribution of authors with the roles and publication speed of authors placed in the centre of the LIS network.

147 (3.8%) of the articles were single-authored and written by central authors, while 748 (19.6%) of them had central co-authors. Central authors' papers were published in a short time compared to the other authors' papers. Kruskal Wallis test confirms this finding ($H(2)$=146.897, p<0.001, η_H^2=0.038).

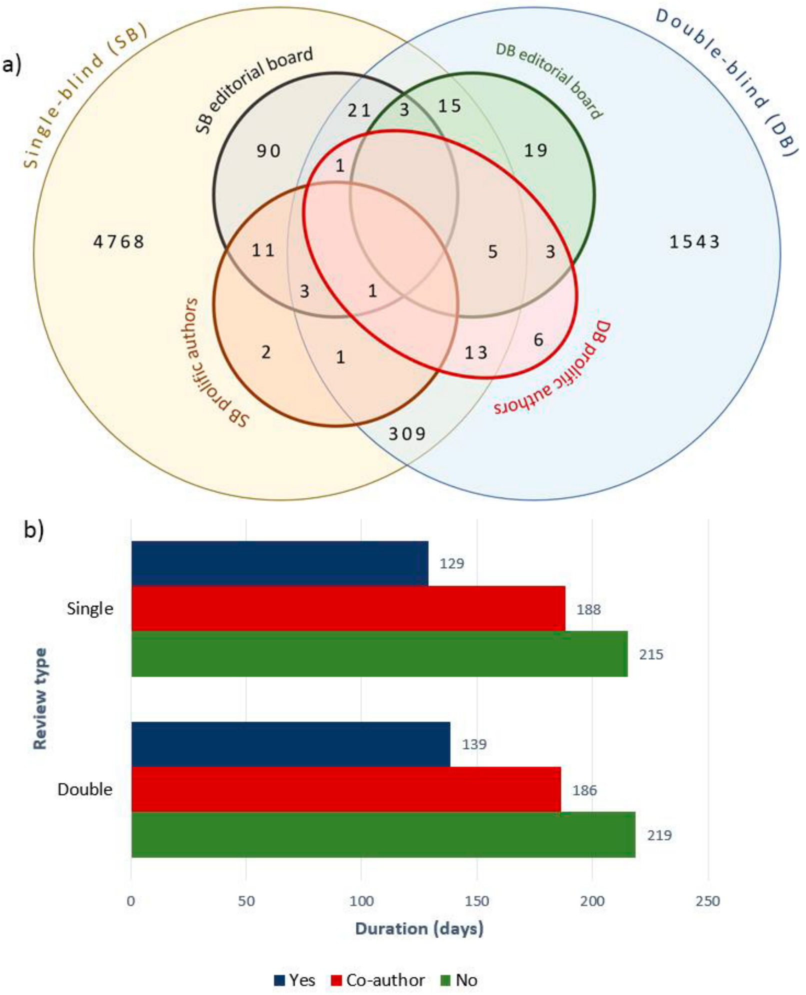


Figure 5. a) Number of unique authors in LIS journals and distribution of their roles, b) Publication speed of articles written by 194 central authors that have two or more roles in LIS

Discussion and Conclusions

According to our results, papers belonging to editorial board members and prolific authors are published faster. This is of course closely related to the experience these people have and the quality of the paper. However, it is also very important in terms of showing the advantage of young researchers who work with these researchers. This study shows that researchers who publish papers with popular and successful researchers have the advantage on the speed of publication.

In addition to the duration advantage of some researchers, this study also shows the importance of sharing the processing dates of research in detail by the publishers. Bilalli et al. (2020) indicated that all dates of review (receive, revision, acceptance, etc.) must be provided by journals. However, the information provided by journals about the review durations is limited especially for single-blind journals. Also, it is not provided as a metadata element in publishers' databases and it requires data mining. To be able to make accurate analyses, this information should be served by publishers and added to databases.

Limitations and Future Studies

This paper seeks author position related factors affecting the publication speed. Many reasons such as the quality of articles, the date of the article submitted, or the workload of editors, can affect the duration of peer review. However, in this study, we aimed to reveal whether there are researchers who have an advantage on publication speed that can be interpreted as Matthew Effect.

The editorial board members and prolific authors have high scientific levels and qualifications, and the outputs produced by these authors have a significant scientific level. It is expected that the papers written by prestigious authors have good quality and so, they are accepted faster than the papers written by others. However, this study is not about the quality of the papers. It aims to reveal the current practices of journals and to present the duration differences in publication processes of articles written by known and unknown authors. Furthermore, the paper is limited to only six journals in the LIS field, but, as indicated in the Methodology part, the volumes of the journals are not the same. It means the workloads of editors are not equal. On the other hand, the audiences of the journals vary. While some journals can be considered as pure library and information science journals such as *ASLIB* or *OIR*, some journals (*SCIM* or *JASIST*) have more authors from different disciplines. All these factors make the comparison harder. Therefore, more investigations are needed to understand all the factors affecting the duration of publications. This research constitutes the first stage of large-scale research.

It is planned to broaden the research by covering the country of affiliation of researchers, country groups, and collaboration statistics. Conducting a multidimensional statistical analysis is the next planned step of this paper. Besides, an author-level analysis covering more journals will be conducted to understand the duration changes for individual researchers. This will provide us with an opportunity to show all the factors affecting publication speed.

References

Bedeian, A. G., Van Fleet, D. D., & Hyman, H. H. (2009). Scientific Achievement and Editorial Board Membership. *Organizational Research Methods*, *12*(2), 211–238. https://doi.org/10.1177/1094428107309312

Bilalli, B., Munir, R. F., & Abelló, A. (2020). A framework for assessing the peer review duration of journals: Case study in computer science. *Scientometrics*. https://doi.org/10.1007/s11192-020-03742-9

Blank, R. M. (1991). The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review. *The American Economic Review*, *81*(5), 1041–1067.

Budden, A. E., Tregenza, T., Aarssen, L. W., Koricheva, J., Leimu, R., & Lortie, C. J. (2008). Double-blind review favours increased representation of female authors. *Trends in Ecology & Evolution*, *23*(1), 4–6. https://doi.org/10.1016/j.tree.2007.07.008

Cohen, B. H. (2013). Statistical tests for ordinal data. In *Explaining Psychological Statistics*. John Wiley & Sons, Inc.

Cox, D., Gleser, L., Perlman, M., Reid, N., & Roeder, K. (1993). Report of the Ad Hoc Committee on Double-Blind Refereeing. *Statistical Science*, *8*(3), 310–317.

Huisman, J., & Smits, J. (2017). Duration and quality of the peer review process: The author's perspective. *Scientometrics*, *113*(1), 633–650. https://doi.org/10.1007/s11192-017-2310-5

Larivière, V. (2019). Resignation of the editorial board of the Journal of Informetrics. *International Society for Scientometrics and Informetrics*. https://www.issi-society.org/blog/posts/2019/january/resignation-of-the-editorial-board-of-the-journal-of-informetrics/

McCormack, N. (2009). Peer Review and Legal Publishing: What Law Librarians Need to Know about open, Single-Blind, and Double-Blind Reviewing. *Law Library Journal*, *101*(1), 59–70.

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Okike, K., Hug, K. T., Kocher, M. S., & Leopold, S. S. (2016). Single-blind vs Double-blind Peer Review in the Setting of Author Prestige. *JAMA*, *316*(12), 1315–1316. https://doi.org/10.1001/jama.2016.11014

Pardeck, J. T., & Meinert, R. G. (1999). Scholarly Achievements of the Social Work Editorial Board and Consulting Editors: A Commentary. *Research on Social Work Practice*, *9*(1), 86–91. https://doi.org/10.1177/104973159900900107

Snodgrass, R. (2006). Single- versus double-blind reviewing: An analysis of the literature. *ACM SIGMOD Record*, *35*(3), 8–21. https://doi.org/10.1145/1168092.1168094

Sun, M., Danfa, J. B., & Teplitskiy, M. (2021). Does double-blind peer review reduce bias? Evidence from a top computer science conference. *ArXiv:2101.02701 [Cs, Econ, q-Fin]*. http://arxiv.org/abs/2101.02701

Taşkın, A. (2021). *Ataskin/article_date* [Python]. https://github.com/ataskin/article_date (Original work published 2021)

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, *114*(48), 12708–12713. https://doi.org/10.1073/pnas.1707323114

Zhang, T., Shi, J., & Situ, L. (2021). The correlation between author-editorial cooperation and the author's publications in journals. *Journal of Informetrics*, *15*(1), 101123. https://doi.org/10.1016/j.joi.2020.101123

Research managers as data experts? Task areas and competency profiles in IT-based research reporting in Germany

Christoph Thiedig[1], Stefan Schelske[2] and Sabrina Petersohn[3]

[1]*thiedig@dzhw.eu*
German Centre for Higher Education Research and Science Studies, Schützenstraße 6a, 10117 Berlin (Germany)

[2]*schelske@dzhw.eu*
German Centre for Higher Education Research and Science Studies, Schützenstraße 6a, 10117 Berlin (Germany)

[3]*petersohn@dzhw.eu*
German Centre for Higher Education Research and Science Studies, Schützenstraße 6a, 10117 Berlin (Germany)

Abstract

Current Research Information Systems (CRIS) form an increasingly important part of the digital infrastructure of Higher Education and research institutions. The persons operating these infrastructures and making use of them in research reporting and assessment have so far rarely been studied. In this paper, we report preliminary results of a job advertisement analysis on the task and competency profiles of positions in IT-based research reporting in Germany. We find distinct profiles for system implementation, project coordination, and research reporting, suggesting a division of labor between more traditional research management functions and research information management, especially pertaining to IT and data expertise as well as strategic tasks.

Introduction

The information needs of university and science policy as well as research assessment are addressed by an increasingly complex and differentiated digital infrastructure. So called Current Research Information Systems (CRIS) form part of this infrastructure. They are integrated databases or information systems which can be used both at the national and institutional level to collect, store, process, integrate, exchange, present and analyze research information (RI). RI comprises metadata about research activities such as persons as authors or project investigators, research projects, publications and other forms of output, awards, patents or research infrastructures (Ebert et al., 2015; Petersohn, Biesenbender & Thiedig, 2020; Zhang & Sivertsen, 2020). CRIS serve many different purposes such as documenting, reporting and monitoring research activities, providing information for quality control and funding allocation procedures or showcasing research to the public (Biesenbender, Petersohn & Thiedig, 2019; Sivertsen, 2019a). More and more research organizations adopt institutional CRIS solutions (Ribeiro et al., 2016, p. 11) due to the richer informational yield of the varied and integrated types of research information collected and processed in CRIS as opposed to common institutional repositories or multidisciplinary citation databases (Sivertsen, 2019b).

Against this backdrop, it seems necessary to look beyond system features and use cases of CRIS and focus on the persons operating CRIS and making use of it in research reporting and assessment instead. The task area is complex, ranging from “frontend” services in training and supporting CRIS users, increasing CRIS acceptance and compliance as well as to assuring data quality. It may be organized centrally or decentrally with responsibility being shared between organizational units such as the library or research administration (Fondermann & van der Togt, 2017; Kaltenbrunner & De Rijcke, 2017; Simons et al., 2017). Knowledge about task profiles, responsibilities as well as underlying competencies and qualifications needed to perform this potentially new specialized area in research management is lacking, with the exception of first efforts undertaken by Blümel et al. (2018) and Dvořák and de Castro (2019). In a current research project, we aim at closing this gap by investigating task profiles, responsibilities and competencies for digitally supported research reporting. In this paper, we report preliminary results of a job advertisement analysis on the task and competency profiles as well as other characteristics of the advertised positions using a trial data set. We further briefly introduce a comprehensive data set of job advertisements that will be the focus of our upcoming research.

Data and methods

In a first step, a comprehensive list of vacancy platforms for the German job market using a list provided by the University of Passau[i] was compiled, which we extended via additional desk and literature research. The relevancy of 229 platforms was assessed based on the search term "research information" (thus also covering mentions of CRIS) as well as the platforms' descriptions. We compiled a trial data set for preliminary analysis from five of these platforms (academics.de, bund.de, interamt.de, jobturbo.de, stepstone.de). Information on vacancies were retrieved based on search terms such as "research reporting", "research documentation", "research service", and "research evaluation", among others. The search results were manually assessed and compiled into an initial trial data set consisting of 32 job advertisements, covering the years 2015 to 2020.

In order to get a more complete picture of the developments of the job market in the realm of IT-based research reporting in Germany, we additionally aimed for the compilation of a comprehensive data set, reaching back to the year 2005, when major CRIS developers first gained traction (Atira, n. d.; Weiss, 2011). Based on our platform assessment and first full-text job offers retrieved as examples, a database query for the creation of a full dataset of job vacancies was developed. Platforms likely to have their own databases were identified and approached with access requests to their full text databases. A corresponding data set was acquired in mid-April of 2021. It contains information on ca. 1100 job openings covering the years 2005 to 2020. We briefly provide first descriptive analyses below.

Analyses of the trial data set was conducted using MaxQDA 2020. A category system was developed covering the following main aspects of the job advertisements: tasks, qualifications, and competencies (understood here as parts of an overarching professional competency; as "…coherent cluster[s] of knowledge, skills and attitudes which can be utilized in real performance contexts", Mulder, 2014, p. 111), as well the type of research institution, year of publication, wage group and further contract of employment characteristics. For example, 226 tasks were coded by 77 inductive categories, such as "development of university-wide reporting and data management", "planning of target group-oriented science communication", etc., which in turn are assigned to one of nine overarching categories - in this case "strategic tasks". While the category system has undergone an intersubjective validation process, no strict intercoder reliability procedure has been conducted yet due to the exploratory nature of the initial analyses and the limited generalizability of the trial data set.

The data were analyzed using qualitative and quantitative content analysis (Mayring, 2001) by drawing on the mixed methods functionalities of MaxQDA. Additional descriptive analyses, correlations and mean comparisons (ANOVAs) were conducted using SPSS and Stata.

Findings

In terms of the characteristics of job vacancies within the realm of IT-based research reporting and CRIS development, we find that most offered contracts are for a limited-term position at a public university requiring a master's degree or equivalent. For IT specialist positions, degree requirements are more diverse and more frequently take previous work experiences into account. Further positions on the "periphery" of operational research information management include department leaders, administrators, and library staff.

Based on the characteristics of the advertisements included in the trial data set, we identify three distinct job profiles: system implementation (n=13), project coordination (n=8) and research reporting (n=9). The profiles are described below, based on our preliminary analysis of tasks (see Figure 1) and competencies (see Figure 2).

System implementation positions are overwhelmingly tasked with the implementation of a CRIS. As such, they are mostly located in the central administration of the institutions. They are characterized by a high number of IT- and data-related tasks ("preparation of research information", "configuration of interfaces") and corresponding competencies ("data visualization skills", "experiences with server environments") as well as frequent and diverse communicative tasks ("user support", "organizing and carrying out events and seminars"). They appear to have almost no association with strategic tasks on an institutional level.

Project coordination positions are further detached from the operational tasks of CRIS implementation, taking a coordinative role and responsibility for both further development and use of a CRIS, sometimes as part of a broader set of responsibilities including, for example, the coordination of external project partners and funding acquisition. Like system implementation positions, they are almost exclusively located in the institution's central administration. "Knowledge of science and university" (for example "knowledge of the research funding landscape" or "knowledge about science communication") are competencies often required for such positions. Department leaders are found in this category. Involvement in the development of reporting systems and research planning is moderate, suggesting that research reporting is a predominantly cross-departmental tasks and thus less the subject of project or department coordination. While a similar amount of IT-related tasks is found in both profiles presented so far, competency requirements pertaining to data expertise and IT systems are significantly more frequent in the former.

In contrast, staff in *research reporting* positions is responsible for developing and extending institutional reporting and service capacities and for strategic purposes such as fulfilling external reporting obligations, research planning, benchmarking or supporting decision-making processes. Positions involving library tasks, for instance "entering data in repositories" or "management of open access funds", are also found in this category. These positions are located in different organizational units relevant to research reporting: institutional leadership, central administration, and on faculty or, in case of extramural research institutions, divisional management level. Involvement in CRIS development, if mentioned at all, is restricted to specific reporting purposes. These positions are more often tasked with data preparation and analyses compared to the other two job profiles. They also more frequently require generic competencies, such as professional communication (e. g. "negotiation skills") and a professional way of working (e. g. "initiative and assertiveness").

Thus, while the differences between the system implementation and project coordination profiles can be said to be of degree and less of kind, the results indicate very distinct job profiles for research reporting on the one hand and CRIS implementation and development on the other.

First analyses of the comprehensive data set indicate a "core" of about 110 positions directly related to research reporting, research information or research documentation, 60 of which explicitly mention a CRIS – a term that first appears in 2009. Another 500 positions are potentially relevant and will need to be assessed further. Perhaps unsurprisingly, the majority of the advertised "core" positions are located at universities (65 percent) and extra-university research institutes (24 percent), most often in central administration (66 percent), with the other positions distributed evenly among the levels of institutional and decentral leadership as well as the library.

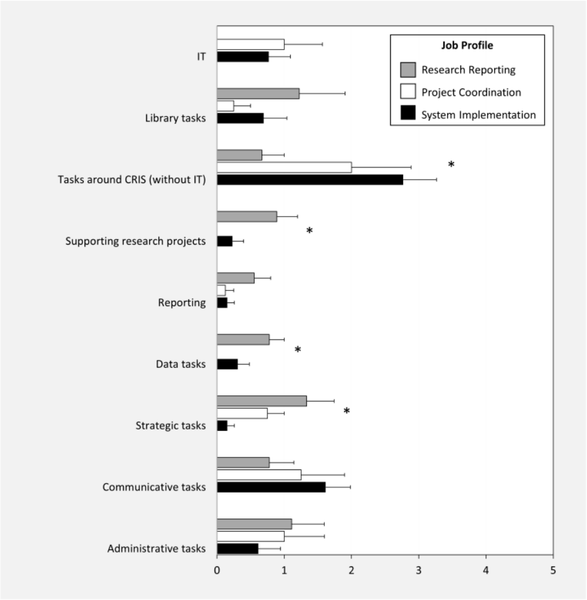


Figure 1. Mean number of tasks by job profile (total number = 226). Error bars show standard errors of the mean. Asterisks highlight significant differences between job profiles within that area of tasks (p < .05, N = 32).

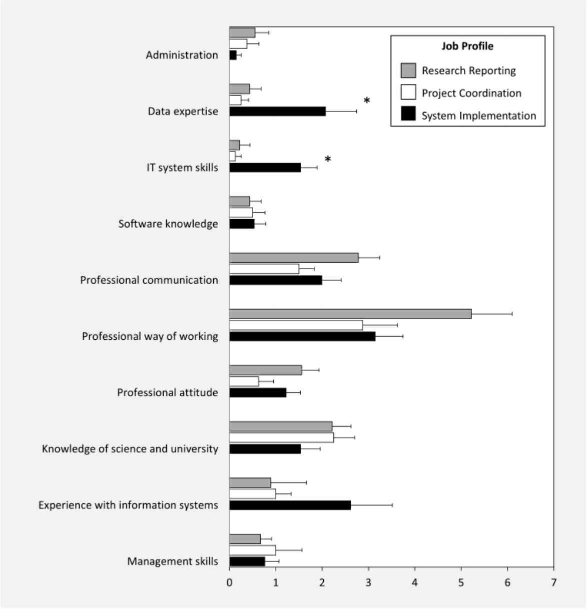


Figure 2. Mean number of required competencies by job profile (total number = 445). Error bars show standard errors of the mean. Asterisks highlight significant differences between job profiles within that area of competence (p < .05, N = 32).

Discussion and conclusion

Digitally supported research reporting is an increasingly relevant task area across all types of research institutions in Germany, and the implementation, development and use of CRIS constitute both benefits and major challenges (Biesenbender & Hornbostel, 2016). Employees are expected to perform a wide range of tasks both in "frontend" and "backend" services. Our study reveals corresponding competency requirements: Personnel responsible for CRIS implementation not only needs to be well versed in project and stakeholder management, but also requires IT and data expertise not commonly requested of other research management positions. The preliminary results thus indicate a rather clear division of labor between these job profiles. Given the exploratory nature of the analyses and the limitations of its underlying data set, further analyses are needed. The comprehensive data set offers several promising avenues for research, allowing for more differentiated analyses of task areas and competencies and the subsequent identification of corresponding job profiles over time, by type of institution, and by type of CRIS (e.g. commercial product vs. in-house development). Analyses on the development of IT and data competency requirements will further gain insights into a common challenge that German research institutions are currently facing: the recruitment and retention of dedicated IT experts. Analytically, the application of competency models from related areas of research support provided by research management and administration, librarians and information professionals, such as the bibliometrics competency model (Cox, Gadd, Petersohn & Sbaffi, 2019), is a promising approach that will be adapted to the field of IT-supported research reporting with a rigorous foundation in the latest developments of research in professional competency (Schelske & Thiedig, 2019). Finally, contextualization with the results of an upcoming online survey on IT-based research reporting in German Higher Education and Research organizations will not only yield further insights into the practices, task areas and competency profiles of personnel in IT-based research reporting beyond their formalization in job advertisements, but also into the kinds and amount of use of research information collected in CRIS and other data bases in institutional (decision-making) contexts – a subject that, so far, little is known about for German research institutions (Hillebrandt, 2020).

Acknowledgments

We gratefully acknowledge financial support from the Federal Ministry of Education and Research (BMBF), grant number 01PW18014.

References

Atira (n. d.). *Whitepaper Pure Research Information System Version 4.11.0*. Retrieved April 26, 2020 from: https://radinfo.univie.ac.at/fileadmin/user_upload/radinfo/Pure.whitepaper.4.11.0.pdf.

Biesenbender, S. & Hornbostel, S. (2016). The Research Core Dataset for the German science system: challenges, processes and principles of a contested standardization project. *Scientometrics*, 106(2), 837–847.

Biesenbender, S., Petersohn, S. & Thiedig, C. (2019). Using Current Research Information Systems (CRIS) to showcase national and institutional research (potential): research information systems in the context of Open Science. *Procedia Computer Science*, 146, 142–155.

Blümel, I., Walther, T., Zellmann, C., Hauschke, C., Wartena, C. & Hahn, L. (2018, June 12-15). *FIS-Curriculum – Bedarfe zur Ausbildung künftiger Forschungsinformations-Manager.* Retrieved April 26, 2021 from urn:nbn:de:0290-opus4-32963.

Cox, A., Gadd, E., Petersohn, S. & Sbaffi, L. (2019). Competencies for bibliometrics. *Journal of Librarianship and Information Science*, 51(3), 746–762.

Dvořák, J. & de Castro, P. (2019, October 21). *The Roles and the competencies of CRIS managers*. Retrieved April 26, 2021 from: http://hdl.handle.net/11366/1175.

Ebert, B., Tobias, R., Beucke, D., Bliemeister, A., Friedrichsen, E., Heller, L., Herwig, S., et al. (2015). *Research Information Systems at universities and research institutions—position paper of DINI AG FIS*. Retrieved April 26, 2021 from: https://zenodo.org/record/17491.

Fondermann, P. & van der Togt, P. L. (2017). How Wageningen University and Research Centre managed to influence researchers publishing behaviour towards more quality, impact and visibility. *Procedia Computer Science*, 106, 204–211.

Hillebrandt, M. (2020). Keeping one's shiny Mercedes in the garage: why higher education quantification never really took off in Germany. *Politics and Governance,* 8(2), 48-57.

Kaltenbrunner, W. & De Rijcke, S. (2017). Quantifying 'output' for evaluation: administrative knowledge politics and changing epistemic cultures in Dutch law faculties. *Science and Public Policy*, 44(2), 284–293.

Mayring, P. (2001). Kombination und Integration qualitativer und quantitativer Analyse. *Forum: Qualitative Social Research*, 2(1).

Mulder, M. (2014). Conceptions of professional competence. In S. Billett, C. Harteis & H. Gruber (Eds.), *International Handbook of Research in Professional and Practice-based Learning*, Springer International Handbooks of Education (pp. 107-137). Dordrecht: Springer Netherlands.

Petersohn, S., Biesenbender, S. & Thiedig, C. (2020). Investigating assessment standards in the Netherlands, Italy, and the United Kingdom: challenges for responsible research evaluation. In K. Jakobs (Ed.), *Shaping the Future Through Standardization* (pp. 54–94). Hershey, PA: IGI Global.

Ribeiro, L., Mennielli, M. & de Castro, P. (2016, March). *Final Report -EUNIS EuroCRIS joint survey on CRIS and IR. ERAI (Eunis Research and Analysis Initiative)*. Retrieved April 26, 2021 from: https://www.eunis.org/wp-content/uploads/2016/03/cris-report-ED.pdf.

Schelske, S. & Thiedig, C. (2019, November 20). *New professional roles? Competencies and task profiles in IT-supported research reporting – introduction to the research project "BERTI"*. Retrieved April 26, 2021 from: https://dspacecris.eurocris.org/handle/11366/1229.

Simons, E., Jetten, M., Messelink, M., van Berchum, M., Schoonbrood, H. & Wittenberg, M. (2017). The important role of CRIS's for registering and archiving research fata. The RDS-project at Radboud University (the Netherlands) in cooperation with Data-archive DANS. *Procedia Computer Science*, 106, 321–328.

Sivertsen, G. (2019a). Developing Current Research Information Systems (CRIS) as data sources for studies of research. In W. Glänzel, H. F. Moed, U. Schmoch & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators*, Springer Handbooks (pp. 667–683). Cham: Springer International Publishing.

Sivertsen, G. (2019b). Understanding and evaluating research and scholarly publishing in the social sciences and humanities (SSH). *Data and Information Management*, 3(2), 61–71.

Weiss, R. (2011). *CONVERIS and RoMEO- Information System for Research and Innovation*. Retrieved April 26, 2021 from: http://www.rsp.ac.uk/documents/get-uploaded-file/?file=AVEDAS-CONVERIS%20and%20RoMEO%20V1%20RW.pdf.

Zhang, L. & Sivertsen, G. (2020). The new research assessment reform in China and its implementation. *Scholarly Assessment Reports*, 2(1), 3.

[i] https://www.uni-passau.de/careersup/externe-stellenboersen-nach-branchen/, last accessed on April 26, 2021.

Comparing different implementations of similarity for disparity measures in studies on interdisciplinarity

Bart Thijs[1], Ying Huang[2] and Wolfgang Glänzel[3]

, [1]*bart.thijs@kuleuven.be,* [2]*ying.huang@kuleuven.be,* [3]*wolfgang.glanzel@kuleuven.be*
KU Leuven, ECOOM & Dept MSI, Leuven Belgium

Abstract

The recent literature on interdisciplinarity in quantitative science studies focuses on the development and interpretation of measures of diversity. These measures consider the mutual similarity or distance between disciplines in the applied subject scheme. Most of the studies on this topic are well aware of the importance of the choice of the proper classification scheme as it may lead to large differences in the obtained diversity scores. The central objective of the present study is to investigate the underlying properties of distinct similarity matrices used as starting point in quantifications of disparity. The study screens ten combinations of three different classification schemes and five implementations of citation or reference-based similarities using a version of a cosine similarity. In addition, each of the ten combinations are calculated for nine sliding time windows. The ten combinations are scored on different evaluative criteria both of quantitative and qualitative nature: *stability*, *discriminative power*, *density*, *skewness and deviation* and *ease of calculation*. The study provides the required tools for an informed choice on the appropriate similarity measures in future research on and application of diversity measures. Based on the investigated criteria, the study favors the use of bibliographic coupling on a medium-resolution granularity subject classification.

Introduction

In quantitative science studies of interdisciplinarity, the measurement of diversity takes a prominent place. The concept was proposed by scientometricians in analogy to its application in ecology (e.g., Macarthur, 1965) and still there is a spill-over from the latter field to scientometrics. Thus, the relevant work from ecologists, most notably that by Jost (2006, 2007 and 2009) and Leinster & Cobbold (2012) is often cited in the scientometric literature (e.g. Wang et al., 2015; Zhang et al., 2016), who stated that "*choices of different subject schemes may lead to different diversity results*" and demonstrated this at both the level of individual paper and at the aggregated level of scientific journals. It would be remiss to mention in this context, that the question of the choice of the aggregation level for the measurement of diversity forms part of the subject of a separate study by Huang et al. (2021). This is the reason, why we do deal with this issue in our recent study. Furthermore, we will not focus on the development or the improvement of existing indicators for measurement in the context of IDR, either. The central issue of the present study is laid on the scientometric methods *underlying* the quantification and measurement of diversity in IDR. This issue is twofold; on the one hand, one has to choose the particular (dis-)similarity measure to define the distance (similarity) between individual disciplines and, on the other hand, one has to select the scientometric method for analysing the links leading to the knowledge integrated into the interdisciplinary documents under study. The latter aspect is commonly used in the context of cognitive links that are manifested as citation and/or textual relationships in the published scientific literature. The particular objective – be it document-clustering exercises or network analysis – is in this connection secondary. In the present study we will use bibliographic coupling (BC), co-citation (CC) and direct citation/cross-citation (CRC) and their combinations. But analogously to the mapping-of-science exercises, also a combination with text-based similarities (cf. Glänzel & Thijs, 2011) are possible to improve the results, most notably in fields, where citations in periodicals play a less significant role as, for instance, in the humanities and in several disciplines in the social sciences.

A further aspect emerges when applying scientometric methods to IDR studies; apart from the already mentioned aggregation level, the issue of granularity emerges as well. This refers to the question of at what subject level interdisciplinarity is to be investigated, a fine-grained topic level, the more moderate sub-field or discipline level or the more general view at the larger research areas. As this raises not only technical questions but also conceptual ones, granularity related issues are discussed in detail in a separate study by Gänzel et al., (2021). Nonetheless, we will extend our analysis to three different levels of granularity within the Leuven-Budapest subject-classification scheme to give evidence of the significance of this issue.

In the study, we will analyse a combination of various methodological settings and using a large dynamic document set retrieved from the three main journal editions of Clarivate Analytics Web of Science Core Collection to find an optimum solution for the implementation of similarity measures in studying the aspects of variety and disparity in interdisciplinarity.

Methodology and Data

Data Source

All documents indexed as articles, letters or reviews in the three journal editions (SCIE, SSCI, A&HCI) of the 2006-2018 volumes of the Web of Science Core Collection (WoS) have been extracted from this database. The data set contains more than 24.4 million so-called citable publications of types 'article, letter or review'. References indexed with the publications are processed and matched with both the cited paper and with similar references in other publications. Each publication is labelled in three classification systems.

Classification Schemes

Three different levels of classification, each with a distinct level of granularity are selected. At the most fine-grained level are the subject categories from the Web of Science Core Collection also known as 'Web of Science Categories'. This classification system holds 255 subject areas, and each journal indexed in this database is assigned to one or more of these subject categories. The system is dynamic as categories can be added yearly (*'Audiology & Speech-Language Pathology'*, *'Nanoscience & Nanotechnology'*) or can become obsolete (*'Biology, Miscellaneous'*). Changes in the system are not in retrospect applied to issues of journals already indexed earlier in the database.

For the analysis at the meso-level, we use the 74 disciplines or subfields of the updated Leuven-Budapest scheme (Glänzel & Schubert, 2003; Glänzel, Thijs & Chi, 2016). The scheme applies a hierarchical system on top of the Web of Science categories. This system is less prone to changes as the total number of disciplines remains stable. Of course, dynamics at the level of individual journals are captured through yearly assignments, allowing shifting profiles over time.

The highest level of aggregation comes from the 16 major subject fields from the same Leuven-Budapest scheme.

The choice of a particular classification scheme is in concordance with the fixed framework assumption stated by Rousseau (2019) and requires that the classification scheme used for final disparity or diversity calculations cannot be altered through the deletion or addition of any particular class.

Similarity Measures

This study proceeds from a vector space model. One of the applicable native similarity measures in a vector space is the uses the cosine measure, which is actually the cosine of the angle between the reference/citation vectors representing the two documents formed by the scalar product of the two vectors divided by the product of their lengths (see equation below). Moreover, in a Boolean vector space as defined by bibliographic coupling and co-citations between individual documents, the cosine measure becomes simply identical with Salton's measure, i.e., the number of joint references/citations divided by the geometric mean of the total number of references/citations of the two documents. Note that this approach applies to individual documents, while cross-citation measures, although methodologically based on direct citations, are applied to document sets and, therefore, require a different approach, particularly for two reasons: The underlying (direct) citation links are not symmetric and unit self-transactions often result in matrices with dominant main diagonals. We will use the formula proposed by Lin et al. (2015) to avoid possible biases caused by this effect (see second equation below).

Based on the choice of underlying citation link a different matrix is subject to this calculation. At each level of classification, a vector is created for the distinct subject classes, and a scalar product is taken between the vectors, here denoted by a and b.

$$similarity(\boldsymbol{a}, \boldsymbol{b}) = \frac{\sum_i^n a_i \, b_i}{\sqrt{\sum_i^n a_i^2} \sqrt{\sum_i^n b_i^2}}$$

1. *Bibliographic Coupling (BC):*

   The elements (i) of the vector representing a (sub)field in the vector space are integer values, which express the number of references to documents in the database. In particular, these elements take the values 1 or 0 according as the document *i* is cited or not. The length of the vector is equal to the number of citable documents in the database.

2. *Indirect Bibliographic Coupling (BC^{ind}):*

   The elements (i) of the vector representing a (sub)field in the vector space are integer values, which express the number of references to all documents assigned to a particular (sub)field in the database. These elements take the values 0, 1, 2, … according to the absolute frequency of the cited references. The length of the vector is equal to the number of (sub)fields in the applied classification scheme. This approach is referred to as *indirect bibliographic coupling* as it links (sub)fields through the field classification of the shared references. This is an indirect link as it allows two (sub)fields that do not refer to the same publication at all to be linked and it introduces a *partial* aggregation, i.e., on the side of the references. The implementation outlined here is similar to that proposed by Leydesdorff and Rafolds (2009) for the creation of global science maps.

3. *Co-Citation (3-year citation window) (CC^{3yr}):*

   The elements (i) of the vector representing a (sub)field in the vector space are integer values, which express the number of citations from documents in the database. Only citing documents indexed in the same year as the cited document or in the two subsequent years are taken into consideration. The elements take the values 1 or 0 according as the document *i* has cited the document or not in this citation window. The length of the vector is equal to the number of citing documents in the database in the period.

4. *Co-Citation (Open citation window) (CCO):*

The elements (i) of the vector representing a (sub)field in the vector space are integer values, which express the number of citations from documents in the database. All citing documents, without any restriction with respect to the year of indexing are considered. The length of the vector is equal to the number of citing documents in the database.

5. *Cross-Citation (CRC):*
While BC and CC are based on individual-document relationships, which can be aggregated to any level, in our study to the subject level of different granularity, cross-citation links, as a combination of references and citations derived from direct citations, are defined on document sets. The elements (i) of the vector representing a document set, i.e., a field or discipline, in the vector space are integer values, which express the number of references and citations to all documents assigned to a particular field in the database. The length of the vector is equal to the number of fields or disciplines in the applied classification scheme. As similarity is usually understood as a symmetric relationship, the resulting matrix needs to be symmetrised. The field or class vector contains binary values that denote whether the field is at the citing or cited side of a particular reference. The length of the vector is equal to the number of references in the database. In fact, the vectors are rows in the incidence matrix describing the undirected and unweighted version of the citation network between the subject fields. This implementation of the cosine similarity results in exactly the same measure as the similarity based on the normalised cross-citation matrix (Zhang et al., 2016) which is as follows

$$s_{ij} = \frac{c_{ij} + c_{ji}}{\sqrt{(TC_i + TC_i)(TC_j + TR_j)}} \; (with\, i \neq j)\,,$$

where i and j refer to subject fields, (c_{ij} + c_{ji}) denotes to the number of cross-citations between subjects i and j and *TC* (*TR*) the total number of citations received by (given to) the subject i and j, respectively from (to) all other subject fields. This formula is also used by Wang and Schneider (2020) in their study on the inherent relation among interdisciplinary measures.

Selected combinations

From the 15 possible combinations of granularity levels provided by the Leuven-Budapest classification scheme and the proposed similarity measures, ten have been selected for further analysis. These ten combinations are summarised in Table 1. (Direct) bibliographic coupling is calculated for all three schemes. The two variants of co-citation are applied on the two levels of the Leuven-Budapest scheme as is cross-citation. The indirect version of bibliographic coupling is restricted to the 74 disciplines. As nine publication windows are used for each selected combination, the study is finally based on 90 different datasets. For the complete Leuven-Budapest scheme consult the Appendix and Clarivate Analytics database website (Clarivate Analytics, 2021) for the underlying about 250 WoS categories.

Publication Windows

Nine distinct but overlapping publication windows of three years are used. The first window comprises publications between 2008–2010, the followed by 2009–2011, then 2010–2012, … until the last one 2016–2018. This scheme provides a kind of 'moving average' approach with slightly smoothing the observed trends. As a result, 90 different similarity matrices are used in this study.

Table 1. Selected combinations of classification schemes and similarity measures

Classification Scheme	BC	BC^{2nd}	CC^{3yr}	CCO	CRC
16 Major Subject Field	*BC_16*		*CC^{3yr}_16*	*CCO_16*	*CRC_16*
74 Disciplines	*BC_74*	*BC^{ind}_74*	*CC^{3yr}_74*	*CCO_74*	*CRC_74*
255 Subject Categories	*BC_255*				

Evaluation Criteria

Before we present the properties of the different combinations, it is reasonable to discuss the different criteria that will be taken into consideration when evaluating the appropriateness, applicability and usability of these implementations of the similarity measures at different levels of granularity. Both quantitative as qualitative criteria are relevant in this context.

Stability

The interpretation of this first criterion is two-fold. We interpret stability primarily in terms of dynamic robustness, that is, as stability of the distribution of the indicator over time. This includes that changes in general descriptive statistics remain within reasonable boundaries. Large changes or fluctuations would influence the final scores that will be obtained. Hence, this time-based instability is to the detriment of the validity of the statement with respect to the dynamics of final scores. This also implies that the same baseline and reference standards may be applied to different periods.

On the other hand, as subject fields are dynamically evolving and so are the relationships between the fields, the underlying similarity matrices calculated over different periods must be able to capture and mirror those in an adequate manner. Stability should not evolve into high rigidity.

Discriminative Power

Also, this criterion is twofold as it indicates the degree to which the similarity is able to assign different scores to pairs of different subjects and, on the other hand and as a matter of course, different scores must be associated with differences in the strength of the relationship between subjects. As such, the discriminative power of a particular implementation of the class similarity also relates to the ability to detect the dynamics in the relations between (sub-)fields and should be balanced with the stability.

Density

The similarity matrix between the different classes at any granularity level can be considered as the weighted adjacency matrix of the network. Non-zero elements indicate the presence of a link between the subjects. At the global level, the density of the network indicates the number of present links compared to the number of possible links. In this study, the matrix representing the network should be as complete as possible, i.e., the number of zero-elements needs to be minimized. Given the dynamics of science, we cannot exclude that non-existent links between two disciplines will never emerge. Interdisciplinarity and new emerging topics are the most striking examples for such changes. And, as the similarity between subjects is often used in the denominator of the calculation of the indicator, setting the value to zero would result in infinite indicator values.

Skewness and deviation

These two statistical functions reflect important properties of empirical distributions in general. As such, they describe the shape and the asymmetry of the distribution of the similarities in our case. Large standard deviations of the similarity distributions combined with low skewness support the discriminative power at the level of individual pairs of disciplines or fields.

Ease of Calculation

Last but not least, the "computability" of the indicators remains an important criterion of applicability and replicability. This criterion particularly refers to the amount of data required to calculate the matrix and to the availability of the data.

Results

The first results with the descriptive statistics for the comparison of the ten selected scenarios and two different time frames are presented in Table 2. Both the mean of the similarity and the density of the network is given.

Table 2. Mean similarity and network density for 10 selected combination given for the first and last 3-year time period considered

	Mean Similarity		**Density of the network**	
	2008-2010	2016-2018	2008-2010	2016-2018
BC_16	0.073	0.100	100%	100%
BC_74	0.024	0.033	97%	99%
BC_255	0.051	0.050	14%	22%
BC^{ind}_74	0.054	0.066	100%	100%
CC^{3yr}_16	0.090	0.098	100%	100%
CC^{3yr}_74	0.067	0.073	100%	99%
CCO_16	0.098	0.098	100%	100%
CCO_74	0.034	0.033	100%	99%
CRC_16	0.042	0.044	100%	100%
CRC_74	0.010	0.009	97%	98%

The first striking observation is that nine out of the ten combinations result in a complete or near-to-complete network. Only the density of the network at the lowest level of granularity with values from 14% to 22% is much lower. This means that most links between subject categories are not present in the selected time windows. As mentioned above, such an absence of pairs of categories in the matrix might pose problems in the calculation of indicator values as it results in undefined indicators. Some of the scenarios do not show any evolution in their mean score. This is the case for the open-ended Co-Citation (CCO), for the Bibliographic Coupling at the level with the finest granularity and for the Cross-Citation. This reflects extreme stability of the obtained similarities but neglects possible structural dynamics that might occur. Breaking down the 16 fields in the ECOOM classification system to the 74 disciplines versions lowers the mean scores. A further breakdown of the 255 subject categories creates multiple edge cuts in the network removing the lowest weights. This explains the higher mean value as non-edges are not set to zero and considered missing values.

In a next step, probability density functions (PDF) are calculated for all ninety distribution. These functions specify the probability that a value or observed similarity is within a particular range. More formally, this probability would be the result of the integral of that distribution taken over the specific range. Figures 1, 2 and 3 plot these functions for the ten different combinations.

All ten plots confirm the different patterns presented in Table 2. The first four plots in Figure 1 show the PDF for the bibliographic coupling-based combinations. It is immediately clear that the choice of classification scheme has an enormous impact on the obtained similarity measures.

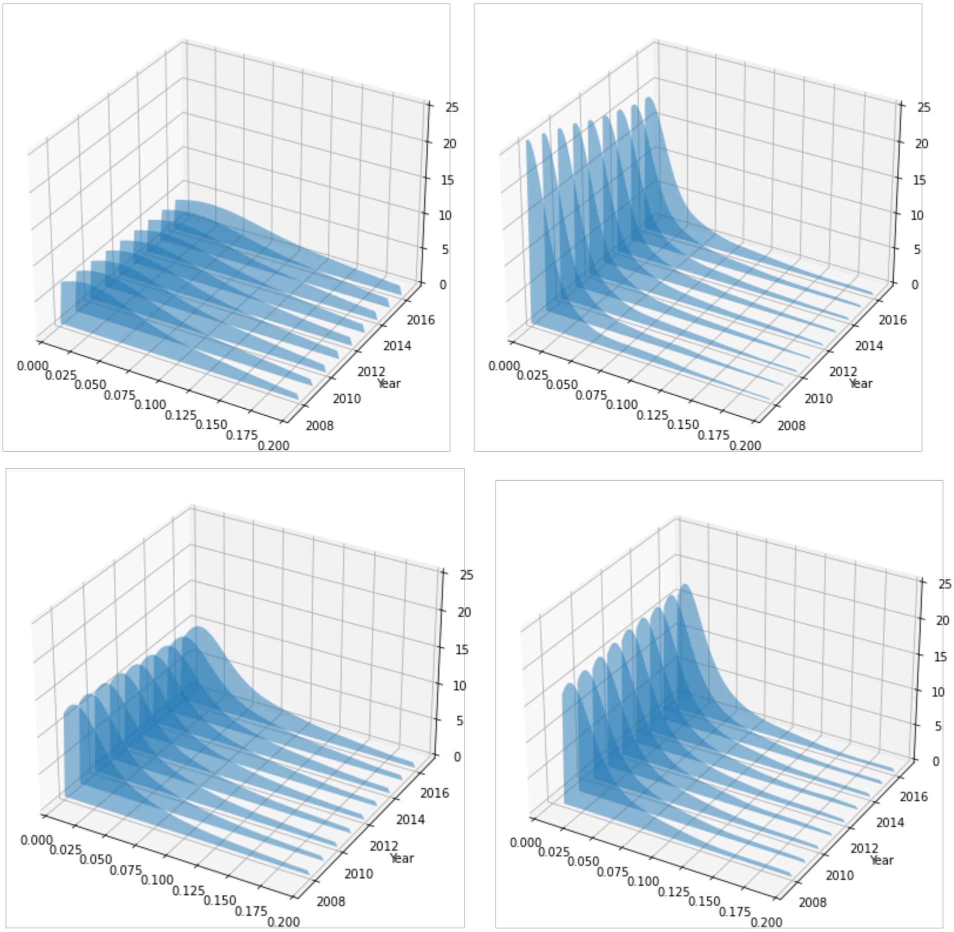


Figure 1. Probability Density Function of the similarities using bibliographic coupling (top left: *BC_16*, top right: *BC_74*, bottom left: *BC*^{ind}*_74*, bottom right: *BC_255*).

The plot in the top-left corner of Figure 1 uses the 16 fields in the ECOOM classification and has a very flat distribution with high standard deviation and low skewness. This is contrasted by the more skewed distribution of the 74 disciplines alternative in the top-right corner. These plots support the capabilities of this combination to capture structural changes. High-rank correlations between the time slices support claims on the stability of this option.

The indirect version of bibliographic coupling on 74 disciplines, bottom-left in Figure1, shows more variation of the scores. The dynamics are retained but less pronounced, and the mean value only increases by 21% compared to 37% in the (direct) BC. The standard deviation starts with a higher value and the skewness is lower. This approach, however, suffers from a more complex calculation.

The Bibliographic coupling with the about 250 subject categories has a similar pattern of the probability density functions. The pattern is only calculated for the existing, non-zero values.

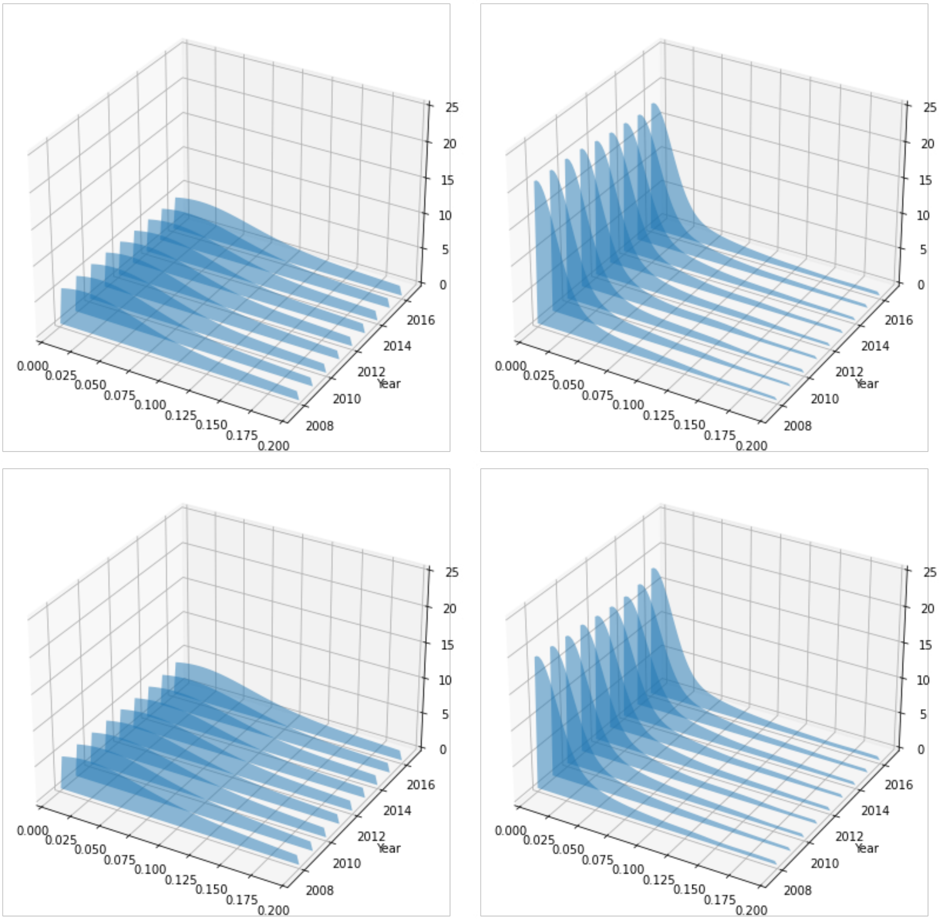


Figure 2. Probability Density Function of the similarities using Co-Citation (top left: *CC*^{3yr}*_16*, top right: *CC*^{3yr}*_74*, bottom left: *CCO_16*, bottom right: *CCO_74*.

Figure 2 holds the four alternative versions of the Co-Citation based approach. The plots on the top are based on a three-year citation window, while the bottom row holds the approaches with an open citation window. The differences in flatness between left and right in Figure 2 is analogous to the top row in Figure 1. It is based on the distinction between the 16 and 74 classes in the applied classification scheme. The effect of the applied citation window is marginal. All four plots show a quite rigid pattern of the similarity distributions and underpin the lack of the ability to capture structural changes over time.

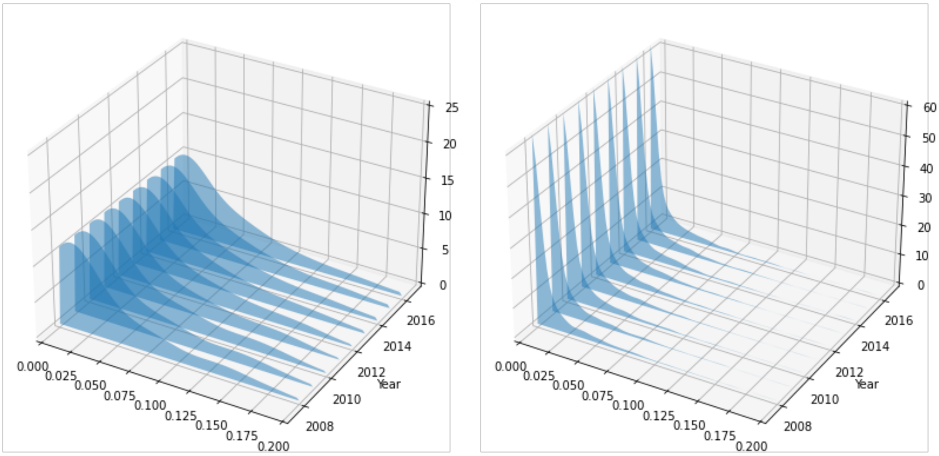


Figure 3. Probability Density Function of the similarities using Cross-Citation (left: *CRC_16*, right: *CRC_74*)

The last two plots are presented in Figure 3. These distributions are more skewed than any of their counterparts with the same underlying classification scheme. In fact, the plot on the right-hand side has an extreme pattern with a steep spike close to zero indicating that almost all obtained similarity scores are very small. This distorts the discriminative power.

Table 3. Summary of the performance of the 10 combinations on five evaluative criteria

	Stability	*Density*	*Discriminative power*	*Skewness & Std Dev*	*Ease of calculation*
BC_16	X	X			X
BC_74	X	X	X	X	X
BC_255	X		X	X	
BC$^{\text{ind}}$_74	X	X	X	X	
CC$^{\text{3yr}}$_16		X			X
CC$^{\text{3yr}}$_74		X	X	X	X
CCO_16		X			X
CCO_74		X	X	X	X
CRC_16		X		X	
CRC_74		X			

Finally, Table 3 provides a summary of the scoring of the ten different scenarios on these five criteria. As mentioned above, these criteria have not only a pure quantitative nature but also require a more qualitative interpretation.

Conclusions

The first observation from the presented analysis relates to the broad range of distributions that are obtained with different combinations of classification schemes and similarity measures. Consequently, this confirms the importance of the proper choice of combination as already raised in the introduction. This choice will have substantial consequences on the final implementation of the disparity and variety measures. Based on the provided material, researchers and users can make the appropriate choice of both classification scheme and similarity measure for their particular application of the diversity score.

Taking all properties and characteristics of each of the combinations into account, this study favors the bibliographic coupling at the level of the 74 disciplines as the most appropriate. This combination provides a nearly complete network. It captures the dynamics in the underlying structure and is still relatively easy to calculate.

Acknowledgments

The research underlying this study is funded by the bibliometric part of the multi-year ECOOM project "Interdisciplinary research & Impact" for the Flemish Government (2019–2023). The authors wish to acknowledge this support.

References

Buckland, M. & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45, 12-19.

Clarivate Analytics (2021), *Web of Science Web Services Expanded. Subject Categories (Ascatype).* Accessible at: http://help.incites.clarivate.com/wosWebServicesExpanded/appendix1Group/ascaCategories.html.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56(3), 357–367.

Glänzel, W., Thijs, B., & Chi, P.S. (2016). The challenges to expand bibliometric studies from periodical literature to monographic literature with a new data source: The book citation index. *Scientometrics*, 109(3), 2165-2179.

Glänzel, W., Thijs, B., & Huang, Y. (2021). Improving the precision of subject assignment for disparity measurement in studies of interdisciplinary research. *Proceedings of the 18th Conference of the International Conference on Scientometrics and Informetrics*, Leuven (virtual meeting), 12–15 July 2021, in this volume.

Huang, Y., Glänzel, W., Thijs, B., & Zhang, L. (2021), A framework for measuring the knowledge diffusion impact of interdisciplinary research. *Proceedings of the 18th Conference of the International Conference on Scientometrics and Informetrics*, Leuven (virtual meeting), 12–15 July 2021, in this volume.

Jost, L. (2006), Entropy and diversity. *Oikos*, 113(2), 363-375.

Jost, L. (2007), Partitioning diversity into independent alpha and beta components. *Ecology*, 88(10), 2427-2439.

Jost, L. (2009), Partitioning diversity into independent alpha and beta components. (Correction note on Jost, 2007). *Ecology*, 90(12), 3593-3593.

Leinster, T., Cobbold, C.A. (2012). Measuring diversity: The importance of species similarity. Ecology, 93(3), 477-489.

Leydesdorff, L. & Rafols, I. (2009). A Global Map of Science Based on the ISI Subject Categories. *Journal of the American Society for Information Science & Technology*, 60(2), 348-362.

Macarthur, R.H. (1965). Patterns of species diversity. *Biological Reviews*, 40(4), 510-33.

Ochiai, A. (1957). Zoogeographical Studies on the Soleoid Fishes Found in Japan and its Neighbouring Regions-III. *Nippon Suisan Gakkaishi, 22*, 522-525.

Rousseau, R. (2019). On the Leydesdorf-Wagner-Bornmann proposal for diversity measurement. *Journal of Informetrics*. 13, 906-907.

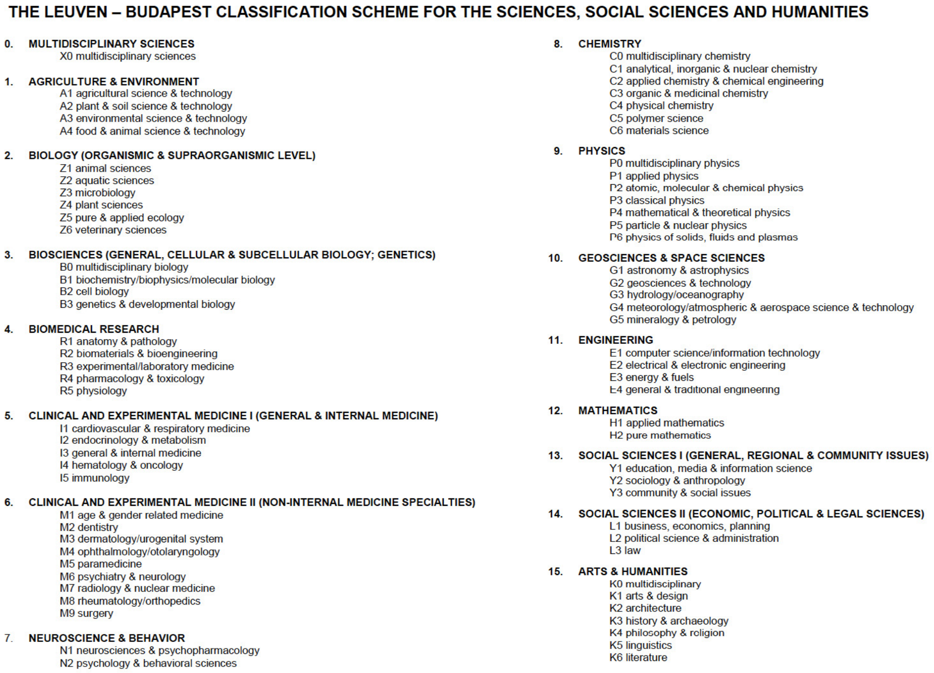
Wang, J., Thijs, B., & Glänzel, W. (2015). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PloS ONE, 10*(5), e0127298.

Wang, Q., & Schneider, J. W. (2020). Consistency and validity of Interdisciplinarity Measures. *Quantitative Science Studies*, 1 (1): 239–263.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: taking similarity between subject fields into account. *Journal of the association for information science and technology*, 67(5), 1257-1265.

Appendix

The revised Leuven-Budapest classification scheme according to Glänzel et al. (2016)



THE LEUVEN – BUDAPEST CLASSIFICATION SCHEME FOR THE SCIENCES, SOCIAL SCIENCES AND HUMANITIES

0. **MULTIDISCIPLINARY SCIENCES**
   X0 multidisciplinary sciences

1. **AGRICULTURE & ENVIRONMENT**
   A1 agricultural science & technology
   A2 plant & soil science & technology
   A3 environmental science & technology
   A4 food & animal science & technology

2. **BIOLOGY (ORGANISMIC & SUPRAORGANISMIC LEVEL)**
   Z1 animal sciences
   Z2 aquatic sciences
   Z3 microbiology
   Z4 plant sciences
   Z5 pure & applied ecology
   Z6 veterinary sciences

3. **BIOSCIENCES (GENERAL, CELLULAR & SUBCELLULAR BIOLOGY; GENETICS)**
   B0 multidisciplinary biology
   B1 biochemistry/biophysics/molecular biology
   B2 cell biology
   B3 genetics & developmental biology

4. **BIOMEDICAL RESEARCH**
   R1 anatomy & pathology
   R2 biomaterials & bioengineering
   R3 experimental/laboratory medicine
   R4 pharmacology & toxicology
   R5 physiology

5. **CLINICAL AND EXPERIMENTAL MEDICINE I (GENERAL & INTERNAL MEDICINE)**
   I1 cardiovascular & respiratory medicine
   I2 endocrinology & metabolism
   I3 general & internal medicine
   I4 hematology & oncology
   I5 immunology

6. **CLINICAL AND EXPERIMENTAL MEDICINE II (NON-INTERNAL MEDICINE SPECIALTIES)**
   M1 age & gender related medicine
   M2 dentistry
   M3 dermatology/urogenital system
   M4 ophthalmology/otolaryngology
   M5 paramedicine
   M6 psychiatry & neurology
   M7 radiology & nuclear medicine
   M8 rheumatology/orthopedics
   M9 surgery

7. **NEUROSCIENCE & BEHAVIOR**
   N1 neurosciences & psychopharmacology
   N2 psychology & behavioral sciences

8. **CHEMISTRY**
   C0 multidisciplinary chemistry
   C1 analytical, inorganic & nuclear chemistry
   C2 applied chemistry & chemical engineering
   C3 organic & medicinal chemistry
   C4 physical chemistry
   C5 polymer science
   C6 materials science

9. **PHYSICS**
   P0 multidisciplinary physics
   P1 applied physics
   P2 atomic, molecular & chemical physics
   P3 classical physics
   P4 mathematical & theoretical physics
   P5 particle & nuclear physics
   P6 physics of solids, fluids and plasmas

10. **GEOSCIENCES & SPACE SCIENCES**
    G1 astronomy & astrophysics
    G2 geosciences & technology
    G3 hydrology/oceanography
    G4 meteorology/atmospheric & aerospace science & technology
    G5 mineralogy & petrology

11. **ENGINEERING**
    E1 computer science/information technology
    E2 electrical & electronic engineering
    E3 energy & fuels
    E4 general & traditional engineering

12. **MATHEMATICS**
    H1 applied mathematics
    H2 pure mathematics

13. **SOCIAL SCIENCES I (GENERAL, REGIONAL & COMMUNITY ISSUES)**
    Y1 education, media & information science
    Y2 sociology & anthropology
    Y3 community & social issues

14. **SOCIAL SCIENCES II (ECONOMIC, POLITICAL & LEGAL SCIENCES)**
    L1 business, economics, planning
    L2 political science & administration
    L3 law

15. **ARTS & HUMANITIES**
    K0 multidisciplinary
    K1 arts & design
    K2 architecture
    K3 history & archaeology
    K4 philosophy & religion
    K5 linguistics
    K6 literature

Artificial Intelligence Research Planning Using Research Cluster Extreme Growth Forecasts: A Retrospective Calibration

Autumn Toney[1], Dewey Murdick[2], Kevin W. Boyack[3] and Richard Klavans[4]

[1] *autumn.toney@georgetown.edu*
Center for Security and Emerging Technology, Washington D.C. (USA)

[2] *dewey.murdick@georgetown.edu*
Center for Security and Emerging Technology, Washington D.C. (USA)

[3] *kboyack@mapofscience.com*
SciTech Strategies Inc., Albuquerque, NM (USA)

[4] *rklavans@mapofscience.com*
SciTech Strategies Inc., Wayne, PA (USA)

Abstract

This work presents a retrospective view into forecasts of artificial intelligence (AI) and machine learning (ML) research areas from 2014 to 2017. With the rapid increase of AI and ML technologies being implemented across a wide range of domains, it is important for planners to understand the evolving landscape of AI/ML research. Specifically, planners need information on which areas of AI/ML research to invest in or divest from. Using an automated structuring of scientific research publications from the Scopus database, subsets of broad research areas are defined as research communities and generated from direct citation links in scholarly literature. We select four main categories of AI/ML research (computer vision, miscellaneous AI/ML, natural language processing, and robotics) and investigate which research communities are predicted to grow or decline using a three-year forecasting model. Our informative forecast model achieves a precision of 0.87 and a recall of 0.48 when predicting the growth of the 154 AI/ML-related research communities in 2014.

Introduction

Artificial intelligence and the underlying machine learning technologies driving it (hereafter abbreviated AI/ML) are widely viewed as disruptive forces in today's economy (Wang, 2020). Research analyzing news article mentions of AI/ML highlights the rapid increase in AI/ML references since 2014, indicating that public awareness of new innovations is also recently expanding (Chuan et al., 2019; Fast & Horvitz, 2017). If you had to make AI/ML research investment decisions in 2014, which specific areas would you have predicted to have exceptional growth? How do you choose which areas to invest in? Do you support the most prestigious researchers in the field or those who have made a recent breakthrough? Given lingering questions around AI/ML safety and its future impact on the human experience, the consequences of these decisions are high (Boström, 2014; Russell, 2019).

We present a model, with a calibrated performance record, to help inform longer-term investment and divestment decisions in emerging technology. This forecasting model, introduced by Klavans et al., provides a predicted growth metric for research areas in all of science using Scopus publication data (Klavans et al., 2020). A similar forecasting approach is commonly used in weather predictions, where one needs to evaluate whether emerging storms will (or will not) cause serious damage in three days. Building on these techniques, our model predicts with comparable levels of accuracy which emerging technologies will (or will not) experience extreme growth in three years.

The data-driven research investment input we present here is useful to research leaders, strategic planners, and program managers who need to take into account fundamental research advances and emerging technology development as part of their job responsibilities. We refer to this group of individuals as planners throughout this paper. Forecasting research area growth is designed to help inform decisions, especially when emerging research areas change rapidly and planners are inundated by experts seeking support of their own agendas. By providing a planner with detailed overviews of research communities, they become equipped with comprehensive knowledge that when used in conjunction with their expertise, leads to effective judgements.

In this work, we introduce a method designed to support these decisions within the area of AI/ML research specifically. Each research investment opportunity can be considered a forming storm cell—yet unclear whether it will grow into a full fledged storm. There are various reasons that any given research area stagnates or declines in growth; not all these reasons are indicators that the research area should be abandoned or have their funding withdrawn. For example, a research area could lose the attention of top researchers when it becomes sufficiently applied to the extent that research questions in that domain are no longer compelling. This may indicate that the research community in question requires a different type of investment to achieve desired outcomes.

Prior research has focused on generating large-scale research community networks, analyzing emerging research areas, and characterizing a specific area of research (Boyack et al., 2020; Boyack & Klavans, 2019; Rotolo et al., 2015; Small et al., 2014). However, to the best of the authors' knowledge, our work is the first retrospective forecasting study, specifically on a detailed topic level. AI/ML presents a relevant and compelling case study for retrospective forecasting because of its quickly evolving nature and its increasing integration into new domains.

Our work focuses explicitly on growth forecasts, and we leave detailed interpretations of the best next steps based on our forecasts to future work. We first outline the methods used to implement our model, summarize the structure of the AI/ML research community in 2014, make predictions about which research communities will experience extreme growth, and then evaluate the three-year forecasts in 2017. The research communities with exceptional growth (those that have become full-fledged storms) are identified, along with the nations, labs, and authors leading in these emerging areas.

Methods

In our work, the concept of a research community (RC) is defined by communication patterns (i.e., citations) between researchers and approximates the community explicitly or tacitly working on defined problems (Kuhn, 1970). Using the Scopus database, an automated structuring of the 2014 research landscape identified more than 90,000 research communities from over 30.9 million indexed publications from 1996-2014 (Klavans & Boyack, 2017). Forecasts about RC growth were computed using data on age, citations, sources, and year of peak growth (see (Klavans et al., 2020) for full details). In order to provide research area investment or divestment recommendations, we label RCs based on their broad research areas, identify AI/ML-related RCs, describe the AI/ML RCs, and interpret growth forecasts.

RC labels

Each RC represents a focused subset of scholarly literature derived from a broad research area (e.g., Mathematics). We define broad research areas by grouping related RC documents based on term similarity and labeling the highest level classification based on dominant journal counts using the UCSD science map categories. We label 12 broad research areas: biology, brain science, chemistry, computer science, earth science, engineering, health science, humanities, infectious diseases, medicine, physics/mathematics, and social science. This initial labeling allows us to understand the broad areas of research where AI/ML-related RCs are most active.

AI/ML RCs

To identify AI/ML-related RCs, we first classify AI/ML-related articles in the Scopus database. We apply a SciBERT predictive model trained on arXiv articles (abstract and titles) on the Scopus database (Beltagy et al., 2019; Dunham et al., 2020). SciBERT is a pre-trained language model derived from BERT (Devlin et al., 2018) and is specifically trained on scientific literature text data to support natural language processing tasks, such as classification, in the scholarly literature domain. Dunham et al. implement the SciBERT model on arXiv paper titles and abstracts and train the classifier to identify AI/ML-related articles. Articles published on arXiv are assigned research area labels by their authors. If an article has at least one label of cs.AI, cs.LG, cs.MA, or stat.ML Dunham et al. classified the article as AI/ML; all other research area labels were considered to be non-AI/ML (Dunham et al., 2020).

Dunham et al.'s classifier assigns an AI/ML label, as well as three additional labels for subsets of AI/ML research: computer vision (CV), natural language processing (NLP), and robotics (RO). However, the classifier is not restricted to assigning only one label—an article can be classified as multiple subsets of AI/ML (e.g., CV and RO). We consider an RC to be AI/ML-related if it contains 50% or more AI-related articles. Additionally, if an AI/ML-related RC has 25% or more CV-, NLP-, or RO-related articles we label that RC with the dominant subset of AI/ML; Figure 1 displays a formal diagram of how AI/ML-related RC assignments are made.

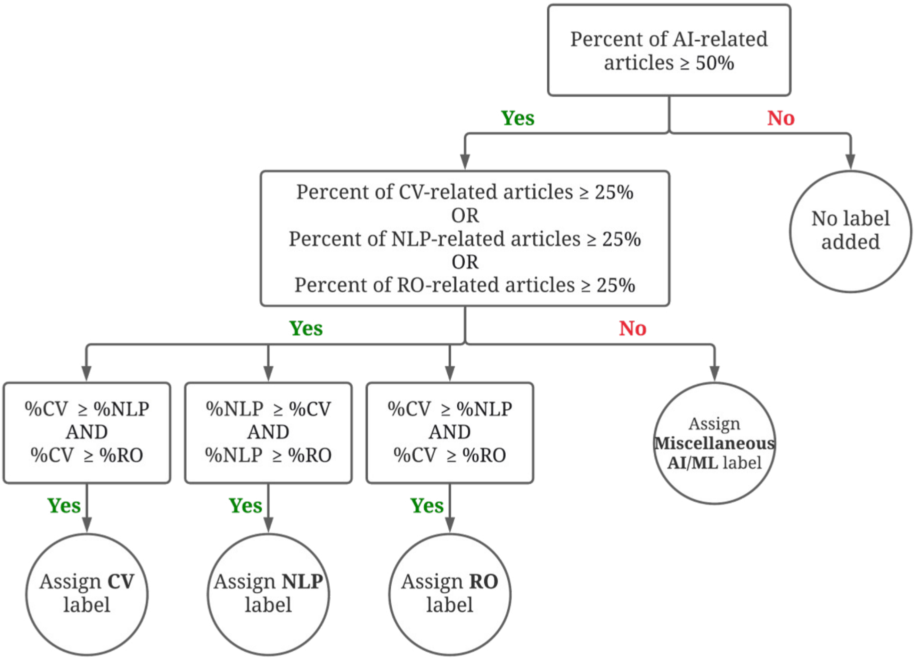

Percent of AI-related articles ≥ 50%
Yes
No
Percent of CV-related articles ≥ 25%
OR
Percent of NLP-related articles ≥ 25%
OR
Percent of RO-related articles ≥ 25%
No label added
Yes
No
%CV ≥ %NLP AND %CV ≥ %RO
%NLP ≥ %CV AND %NLP ≥ %RO
%CV ≥ %NLP AND %CV ≥ %RO
Assign Miscellaneous AI/ML label
Yes
Yes
Yes
Assign CV label
Assign NLP label
Assign RO label

Figure 1. Diagram of AI/ML-related RC assignment.

Interpreting forecast output

Each RC has a predicted growth metric generated by the forecast model, as defined in (Klavans et al., 2020), where an 8% RC annualized growth over three years is considered extreme growth. No additional classifications of growth percentages are labeled. We consider annualized growth percentages that are less than 8% and greater than 0% to be standard growth and growth percentages less than 0% to be declining growth. Table 1 displays the annualized growth percent ranges for each label of growth we reference in the following sections.

Table 1. Growth type labels and characteristics

Growth label	*Annualized growth range (%)*	*Number of AI/ML RCs*
Declining	[-∞, 0)	79
Standard	[0,8)	60
Extreme	[8, ∞)	15

AI RC context

In 2014, 154 research communities were relevant to AI/ML. Of these, 90% of the AI/ML research publication activity was shared within the computer science (CS) literature. Figure 2 displays the AI/ML-related RCs in the context of all of science. The subfigure in the top left corner is the graph of the automated structuring of all RCs in the Scopus database from Klavans et al. (2020), where each dot represents and RC. The location of RCs in Figure 2 is determined by RC citation links, meaning that RCs that have many citation links in common will be closer together. The size of a given RC in Figure 2 corresponds to the number of papers it has in 2014 and the color corresponds to the RC's broad research area label. The main figure highlights AI/ML RCs in the context of all of science, which shows that the majority of AI/ML RCs are in computer science.

We divide the AI/ML-related RCs into the four subset research areas:

1. computer vision, (82 CS and 6 earth science communities),
2. robotics (21 CS and 1 brain science communities),
3. natural language processing (8 CS communities), and
4. miscellaneous AI/ML explorations (36 CS communities).

Planners need to know what communities were likely to have extreme growth and those most likely to shrink. We summarize research community recommendations, based on growth predictions, in the Results section. Of the 154 AI/ML research communities in 2014, 87% of the extreme growth forecasts were accurately predicted for 2017 (i.e., the precision was 0.87). Furthermore, we were able to (in 2014) correctly predict 48% of all RCs that experienced extreme growth in 2017 (i.e., the recall was 0.48).

We use a common weather prediction measure, the Critical Success Index (CSI) (Schaefer, 1990), where CSI = true positives / (true positives + false positives + false negatives), to measure our ability to predict areas of extreme growth. In this case, our CSI = 0.45 is 1.8 times better than the typical three-day weather forecast (CSI = 0.25). The performance of the forecasting model was not highly optimized in order to avoid overfitting future performance to the past, as unforeseen global events can trigger research growth or decline. For example, the COVID-19 outbreak undoubtedly created brand new RCs which our model would be unable to predict accurately. We include further details on the forecasting model performance scores on a public repository (https://github.com/georgetown-cset/AI-research-planning-retrospective-data).

Results

We present results for the four categories of AI/ML-related RCs: 1) Computer vision, 2) Robotics, 3) Natural language processing, and 4) Miscellaneous AI/ML. For each category, we display the RC growth and decline recommendations, as well as additional metadata about the RCs of note (top five highest contributing countries, institutions, and authors). A growth recommendation for an RC is deemed correct if the RC grows at least 8% per year between its most prolific publication year (2014 or before) and 2017—anything less than this extreme growth rate does not merit the recommendation. A decline recommendation is given when a RC's relative growth is predicted to be negative. Incorrect recommendations are highlighted in Tables 2, 4, 6, and 8.

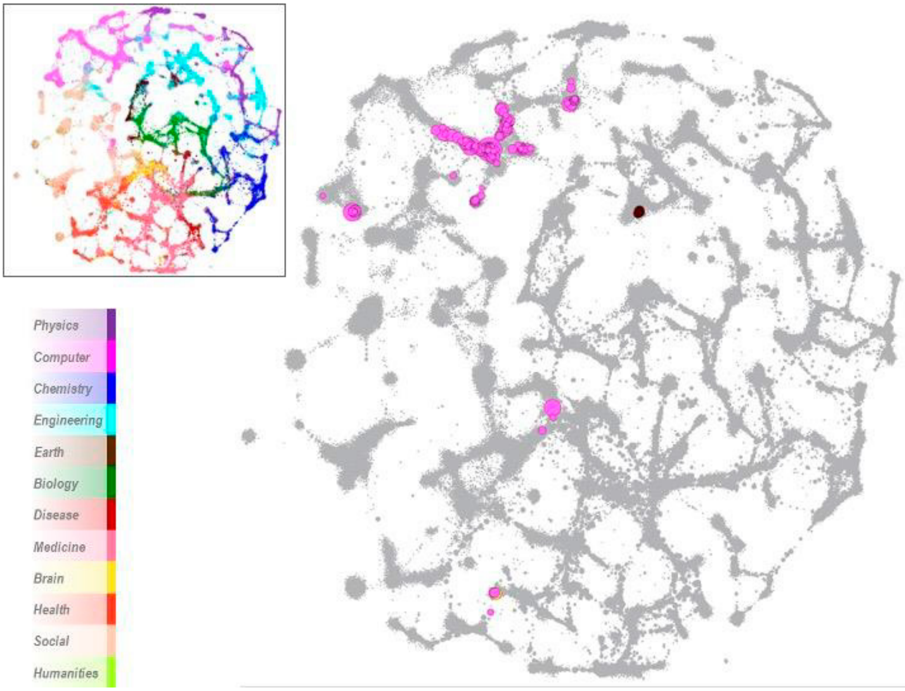


Figure 2. AI/ML-relevant research communities in the context of all of science.

Our results showcase how our forecasting model provides context of RCs and a predictive metric for planners to make a more informed decision about investing in or divesting from a given research area. More details on specific RCs, including information on funding organizations, top cited papers, and predictive features, can be found on a public repository (currently unattached for anonymity).

Computer vision

Computer Vision is an area of computer science and engineering that aims to train computers to gain high-level understanding from images or videos. There are 88 CV RCs, many of which were active and growing in 2014. Of the 88 CV RCs in 2014, 8% are predicted as extreme growth, 39% are predicted as standard growth, and 51% are predicted as declining growth. The top five computer vision RCs with expected extreme growth and expected declining growth are displayed in Table 2. Overall, 90% of the CV RCs are accurately predicted.

Compared to other AI/ML-related RCs (see Tables 4, 6, and 8), computer vision RCs have the highest estimated yearly growth and the "image classification, deep convoluted neural networks" RC has the highest actual yearly growth (103.2%). In the *Research Decline Recommendations*, we incorrectly predict a declining growth (indicated by the highlight in

Table 2) for the "lung cancer, computer-aided diagnosis, computer-aided detection" RC. We consider this a minimal error, as the actual yearly growth was 0.6%. The rest of the *Research Decline Recommendations* are accurate, as well as all of the *Research Growth Recommendations*.

Table 2. Top five computer vision growth and decline recommendations from 2014. Results ordered by estimated three-year annualized growth rate estimated in 2014.

RC#	*Descriptive phrases*	*# Papers (2014)*	*Est. yearly growth (%)*	*Act. yearly growth (%)*
	Research Growth Recommendations			
11244	image classification, deep convolutional neural networks	526	16.1%	103.2%
27987	hash functions, Hamming distance, compact binary codes	132	14.1%	27.5%
8360	scene understanding, scene labeling, convolutional neural networks	350	13.9%	25.9%
17235	image classification, image annotation, image search and retrieval, zero-shot learning	181	13.3%	42.2%
714	visual object tracking, target tracking	770	10.9%	1.7%
	Research Decline Recommendations			
3240	face modeling, facial animation, mouth animation	113	-10.1%	-7.7%
681	video summarization, video retrieval, key frame extraction	177	-9.6%	-5.0%
5826	image enhancement, image processing, anisotropic diffusion	101	-9.6%	-8.7%
7594	image resolution enhancement, image scaling, image interpolation	109	-9.3%	-7.4%
5967	lung cancer, computer-aided diagnosis, computer-aided detection	135	-9.0%	0.6%

Table 3. Top five producers of computer vision research by country, institution and author for the time period 2013-2017.

Country	*Institution (Country)*	*Author (Institution)*
1. China	1. Anna University (India)	1. Zhang, Liangpei (Wuhan University)
2. USA	2. Chinese Academy of Sciences (China)	2. Plaza, Antonio (University of Extremadura)
3. India	3. Tsinghua University (China)	3. Tian, Qi (University of Texas at San Antonio)
4. France	4. Wuhan University (China)	4. Lin, Weisi (Nanyang Technological University)
5. UK	5. Beihang University (China)	5. Shen, Dinggang (University of North Carolina)

We display the top five contributing countries, institutions, and authors across all computer vision RCs in Table 3. These results provide context for the top contributors of the computer vision research area. There is noticeably high contribution from China, and computer vision is the only area of AI/ML where France appears as a top five contributor.

Robotics

Robotics is an area of computer science and engineering research that aims to enable the effective design, construction, operation, and use of robots. In 2014, the robotics research area was relatively smaller (22 RCs) in its application of AI/ML methods than computer vision and had a smaller number of research communities expected to achieve extreme growth in 2017.

Table 4. Top five robotics growth and decline recommendations from 2014. Results ordered by estimated three-year annualized growth rate estimated in 2014.

RC#	*Descriptive phrases*	*# Papers (2014)*	*Est. yearly growth (%)*	*Act. yearly growth (%)*
	Research Growth Recommendations			
21668	soft robotics, continuum robotics, kinematics	119	9.3%	21.4%
	Research Decline Recommendations			
3831	robotic hand, grasp ability, multi-fingered hand	125	-7.3%	-4.1%
4159	self-reconfigurable robot, modular robot, adaptive robots	129	-7.1%	-4.3%
7117	dynamic walking, biped robot, target walking	121	-7.1%	-2.4%
2263	vision-based robotics control, image-based robotics control, mobile robotics	171	-7.0%	-3.8%
3227	robot manipulator, adaptive control, flexible joint robots	126	-6.8%	-4.2%

Of the 22 robotics RCs in 2014, 4% are predicted as extreme growth, 45% are predicted as standard growth, and 50% are predicted as declining growth. Overall, 91% of RO RCs have accurate growth predictions. Table 4 highlights the smaller scale of robotics research activity, as there is only one research cluster with predicted extreme growth. All growth and declining recommendations in Table 4 are predicted correctly.

We display the top five contributing countries, institutions, and authors across all robotics RCs in Table 5. Japan reaches its highest contributor rank in robotics RCs, and robotics is the only area of AI/ML research where South Korea appears as a top five contributor.

Table 5. Top five producers of robotics research by country, institution and author for the time period 2013-2017.

Country	*Institution (Country)*	*Author (Institution)*
1. China	1. Harbin Institute of Tech. (China)	1. Caldwell, Darwin G. (Italian Institute of Technology)
2. USA	2. Italian Institute of Tech. (Italy)	2. Tsagarakis, Nikos G. (Italian Institute of Technology)
3. Japan	3. Carnegie Mellon University (USA)	3. Kubota, Naoyuki (Tokyo Metropolitan University)
4. Germany	4. Mass. Institute of Tech. (USA)	4. Guo, Shuxiang (Kagawa University)
5. S. Korea	5. Beijing Institute of Tech. (China)	5. Vanderborght, Bram (Vrije Universiteit Brussel)

Natural language processing

Natural Language Processing (NLP) is an area of linguistics and computer science seeking to process and analyze large amounts of text and spoken language data to enable effective interactions between computers and human languages. NLP is the smallest area of AI/ML research with only 8 RCs in 2014. Of these 8 NLP RCs, 13% are predicted as extreme growth, 38% are predicted as standard growth, and 50% are predicted as declining growth. Overall, 88% of all NLP RCs have accurate growth predictions. Table 6 displays the one NLP RC with a growth recommendation and the three NLP RCs with declining growth; all recommendations are predicted correctly.

Table 6. Top five NLP growth and decline recommendations from 2014. Results ordered by estimated three-year annualized growth rate estimated in 2014.

RC#	*Descriptive phrases*	*# Papers (2014)*	*Est. yearly growth (%)*	*Act. yearly growth (%)*
	Research Growth Recommendations			
1417	sentiment analysis, opinion mining, sentiment classification	924	11.1%	16.9%
	Research Decline Recommendations			
4613	computational linguistics, human-computer interaction, spoken dialogue systems	108	-8.6%	-4.5%
4882	speech recognition, spoken queries, spoken term detection	136	-7.9%	-3.0%
1621	text classification, term frequency, text mining	393	-7.1%	-2.9%

In 2014, the number of NLP RCs is comparatively small to other AI/ML-related RCs, but the number of papers in the predicted extreme growth cluster (sentiment analysis) is the highest number of all AI/ML-related RCs with 924 articles.

We display the top five contributing countries, institutions, and authors across all natural language processing RCs in Table 7. NLP is the only research area of AI/ML where a UK university (University of Edinburgh) and a company (IBM) appear in the top five contributing institutions.

Table 7. Top five producers of NLP research by country, institution and author for the time period 2013-2017.

Country	*Institution (Country)*	*Author (Institution)*
1. China	1. Tsinghua University (China)	1. Cambria, Erik (Nanyang Technological University)
2. USA	2. Anna University (India)	2. Griol, David (Universidad Carlos III de Madrid)
3. India	3. IBM (USA)	3. Li, Haizhou (National University of Singapore)
4. Japan	4. University of Edinburgh (UK)	4. Yamagishi, Junichi (University of Edinburgh)
5. UK	5. Carnegie Mellon University (USA)	5. Molina, José M. (Universidad Carlos III de Madrid)

Miscellaneous AI/ML

The remaining 36 AI/ML areas cover a wide range of fundamental and applied research tasks. Of the 36 miscellaneous AI/ML RCs in 2014, 17% are predicted as extreme growth, 33% are predicted as standard growth, and 50% are predicted as declining growth. Overall, 89% of all miscellaneous AI/ML RCs have accurate growth predictions.

The top five miscellaneous AI/ML RCs with expected extreme and declining growth are shown in Table 8. In the Research Decline Recommendations, we incorrectly predict a declining growth (indicated by the highlight in Table 8) for the "music genre classification, music recommendation, music information retrieval, and music emotion recognition" RC. We consider this a minimal error, as the actual yearly growth was 0.3%; the rest of the growth recommendations are predicted correctly.

Table 8. Top five miscellaneous AI/ML growth and decline recommendations from 2014. Results ordered by estimated three-year annualized growth rate estimated in 2014.

RC#	*Descriptive phrases*	*# Papers (2014)*	*Est. yearly growth (%)*	*Act. yearly growth (%)*
	Research Growth Recommendations			
32953	extreme learning machine, neural network, incremental extreme learning machine	217	12.9%	18.1%
3437	decision making, intuitionistic fuzzy sets, multiple attribute decision makers	439	12.4%	8.4%
4155	activity recognition, fall detection, wearable sensors	594	10.5%	16.3%
10520	transfer learning, domain adaptation, unsupervised learning	206	9.8%	11.7%
16529	multilabel classification, multi-label learning	179	9.0%	12.3%
	Research Decline Recommendations			
4455	self-organizing map, neural network, topology learning	123	-9.7%	-8.4%
3125	feature selection, cancer classification, gene selection	217	-9.1%	-3.3%
5152	fuzzy rule-based systems, membership functions, genetic algorithms	101	-6.8%	-3.2%
6616	graph matching, pattern recognition, graph classification	121	-6.6%	-2.0%
4335	music genre classification, music recommendation, music information retrieval, music emotion recognition	164	-5.4%	0.3%

We display the top five contributing countries, institutions, and authors across all miscellaneous AI/ML RCs in Table 9. Similar to computer vision results (Table 3), the top five highest contributing institutions are located in China and India, even though the USA is listed above India in the highest contributing countries.

Table 9. Top five producers of miscellaneous AI/ML research by country, institution and author for the time period 2013-2017.

Country	*Institution (Country)*	*Author (Institution)*
1. China	1. Anna University (India)	1. Zhang, Liangpei (Wuhan University)
2. USA	2. Tsinghua University (China)	2. Plaza, Antonio (University of Extremadura)
3. India	3. Chinese Academy of Sciences (China)	3. Caldwell, Darwin G. (Italian Institute of Technology)
4. Japan	4. Harbin Institute of Tech. (China)	4. Tian, Qi (University of Texas at San Antonio)
5. Germany	5. Beihang University (China)	5. Lin, Weisi (Nanyang Technological University)

Conclusion

We have developed a prototype system to forecast growth of research communities in a large-scale highly detailed model of science. This prototype system offers a useful component of a strategic analysis capability. Currently, when planners are asked to make decisions about where to invest their resources in research and development, they lack the global perspective of scientific research as a whole and the estimated growth or decline of specific research areas of interest. Our work provides a unique view into research communities by including growth predictions and metadata about research communities of interest.

We focused our analysis on AI/ML-related research communities, since AI/ML research production is rapidly growing and integrated in many research and applied domains. Our forecasting approach achieved a precision of 0.87 and a recall of 0.48 when predicting the three-year growth of the 154 AI/ML-related research communities in 2014, which is better than most three-day weather forecasts, highlighting its usefulness. We note that these high precision and recall rates may not be maintained in other areas of science and acknowledge that unforeseeable world events are not captured in our predictive capability. Generally, however, our model and forecasting approach have already been evaluated broadly in other fields (Klavans et al., 2020) and have been shown to be a practical and valuable tool for planners who need to make well-informed decisions and currently lack certain important details about the scientific research landscape.

References

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. In *arXiv* [cs.CL]. arXiv. http://arxiv.org/abs/1903.10676

Boström, N. (2014). *Superintelligence: Paths, dangers, strategies*. Google Scholar Google Scholar Digital Library Digital Library.

Boyack, K. W., & Klavans, R. (2019). *Creation and Analysis of Large-Scale Bibliometric Networks*. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), Springer Handbook of Science and Technology Indicators (pp. 187–212). Springer International Publishing.

Boyack, K. W., Smith, C., & Klavans, R. (2020). A detailed open access model of the PubMed literature. *Scientific Data*, 7(1), 408.

Chuan, C.-H., Tsai, W.-H. S., & Cho, S. Y. (2019). Framing Artificial Intelligence in American Newspapers. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 339–344.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *arXiv* [cs.CL]. arXiv. http://arxiv.org/abs/1810.04805

Dunham, J., Melot, J., & Murdick, D. (2020). Identifying the Development and Application of Artificial Intelligence in Scientific Text. In *arXiv* [cs.DL]. arXiv. http://arxiv.org/abs/2002.07143

Fast, E., & Horvitz, E. (2017). Long-Term Trends in the Public Perception of Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). https://ojs.aaai.org/index.php/AAAI/article/view/10635

Klavans, R., and Boyack, K.W. (2017). Research portfolio analysis and topic prominence. *Journal of Informetrics* 11(4), 1158-1174.

Klavans, R., Boyack, K. W., & Murdick, D. A. (2020). A novel approach to predicting exceptional growth in research. *PloS One*, 15(9), e0239177.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*: 2nd Edition. University of Chicago Press.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy*, 44(10), 1827–1843.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin.

Schaefer, J. T. (1990). The Critical Success Index as an Indicator of Warning Skill. *Weather and Forecasting*, 5(4), 570–575.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy*, 43(8), 1450–1467.

Wang, N. (2020, June 26). *Why megacap tech stocks will likely continue leading markets higher*. TheStreet. https://www.thestreet.com/investing/megacap-tech-stocks-will-continue-leading-markets-higher.

The growth of COVID-19 scientific literature: A forecast analysis of different daily time series in specific settings

Daniel Torres-Salinas[1], Nicolas Robinson-Garcia[2], François van Schalkwyk[3], Gabriela F. Nane[2] and Pedro Castillo-Valdivieso[4]

[1] *torressalinas@go.ugr.es*
Information and Communication Studies department, University of Granada, Granada (Spain)

[2] *elrobinster@gmail.com; g.f.nane@tudelft.nl*
Delft Institute of Applied Mathematics, TU Delft, Delft (Netherlands)

[3] *fbvschalkwyk@sun.ac.za*
DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Centre for Research on Evaluation, Science and Technology, Stellenbosch University, Stellenbosch, South Africa

[4] *pacv@ugr.es*
Departamento de Arquitectura y Tecnología de Computadores, University of Granada, Granada (Spain)

Abstract

We present a forecasting analysis on the growth of scientific literature related to COVID-19 expected for 2021. Considering the paramount scientific and financial efforts made by the research community to find solutions to end the COVID-19 pandemic, an unprecedented volume of scientific outputs is being produced. This questions the capacity of scientists, politicians and citizens to maintain infrastructure, digest content and take scientifically informed decisions. A crucial aspect is to make predictions to prepare for such a large corpus of scientific literature. Here we base our predictions on the ARIMA model and use two different data sources: the Dimensions and World Health Organization COVID-19 databases. These two sources have the particularity of including in the metadata information the date in which papers were indexed. We present global predictions, plus predictions in three specific settings: type of access (Open Access), NLM source (PubMed and PMC), and domain-specific repository (SSRN and MedRxiv). We conclude by discussing our findings.

Introduction

The average growth in journal articles published is estimated to be at around 3.0% per annum (Johnson et al., 2018) with an increase to 3.9% between 2006 and 2016. The total for developing countries grew more than twice as fast (about 8.6%) (National Science Board, 2018). Unsurprisingly, and given the scale of scientific output, one of the main research topics within the field of scientometrics has been the study of the growth of scientific literature. Indeed, in the 1960s Derek de Solla Price (1963) had already developed a model of the exponential growth of science in what is considered one of the seminal contributions to the field. Although his contribution was not the first attempt to do model growth (e.g., Coles & Eales, 1917; Hulme, 1923), it reflects the predominant role that the study of bibliometric distributions, dynamics of growth and ageing laws of scientific literature has had in the field.

According to Price's model, there are three distinct phases by which literature increases over time. In the first phase there is a slow increment of publications, followed by an exponential increase, and a third phase in which the curve reaches a saturation point. Since then, different studies have tried to refine his approach, by trying to identify the models which can accurately adjust growth curves for the observed increase in scientific literature (i.e., logistic, power or Gumpertz models)[1]. These studies reflect continued efforts to identify models and distributions which can best adjust to different types of scientific literature. Examples of such studies are those conducted by Egghe and Ravichandra (1992) who observe that Social Sciences literature

[1] An overview is provided by Fernandez-Cano et al. (2004).

appears to fit a Gompertz-S-shaped distribution, while other literatures follow a power law distribution. Similarly, Zhou (2010) analyses the growth of science in China, while Urbizagástegui and Restrepo (2015) apply exponential models to analyse the Brazilian literature.

In this paper we look at scientific growth in exceptional circumstances such as the COVID-19 pandemic. Scientific production on COVID-19 has rocketed in the last year (Torres-Salinas, 2020), reflecting the paramount effort that is being made globally both scientifically and financially to end the global pandemic and to minimize the negative consequences it is having on society. From the scientometric community, efforts have been made to describe the contents of new data sources liberated specifically on the topic of COVID-19 (Colavizza et al., 2020), to compare the coverage of different data sources (Kousha & Thewall, 2020), to analyze the effectiveness of scholarly communication in these pressing times (Homolaket al., 2020; Soltani & Patini, 2020), and its consumption in social media (Colavizza et al., 2020; Thelwall, 2020). The present study is integrated within this stream of literature, building on preliminary findings (Torres-Salinas et al., 2020), and aims to forecast the potential growth of COVID-19 literature to better understand the magnitude of data expected by scientists to cope with the flood of scientific knowledge being produced (Brainard, 2020). We present predictions on the number of COVID-19 publications for 2021. We base our predictions on the Auto-Regressive Moving Average (ARIMA) model and forecast growth in three specific settings. The specific objectives of the paper are summarized as follows:

1. To forecast the growth of publications on COVID-19 in two different databases: Dimensions and WHO.
2. To forecast the growth of publications on COVID-19 in three specific settings to explore the (dis)similarities between them. These are:
    - National Library of Medicine (NLM) databases: Pubmed and PMC
    - Domain-specific scientific repositories: medRxiv and SSRN
    - Type of access to the publications: Open Access and non-Open Access (paywall).

Material and methods

We make use of two different databases: Dimensions and World Health Organization (WHO). The former provides a COVID-19-specific dataset named "*Dimensions COVID-19 publications, datasets and clinical trials*" which is available on FigShare. This dataset contains information on four document types: publications, datasets, clinical trials and grants. In this study, we work only with publications, which have a volume of 168,053 records. The second database is the "*COVID-19 global literature on coronavirus disease*", produced by the WHO. In this case we collected metadata for a total of 113,563 records using the export results option that allows for the downloading of the complete database. These two datasets were collected in December 2020. Like Dimensions, the WHO database contains publications from different sources such as international databases (e.g., Pubmed, Elsevier), databases of international organizations (e.g., WHO COVID-19) and repositories (e.g., medRxiv, SSRN, etc.). One of the characteristics of these two specific databases is that they include for each record the exact date on which publications were indexed. In this sense, we have observed a two-day delay in the indexing dates for the WHO database with respect to Dimensions. This information allows us to establish the daily growth in the number of publications. Table 1 presents a summary of the main characteristics of both databases.

Three different datasets were generated for each database, producing a total of eight time series (Table 2). The first two time series account for the total number of records per day in each database. Two additional time series include the number of published Open Access (OA) and non-OA documents per day. The last four time series refer to the number of documents published by repository. We report predictions of growth for the following repositories: PubMed, PMC, medRxiv and SSRN.

Table 1. Main characteristics of the analysed databases: Dimensions and WHO

	Dimensions	**WHO**
Link	https://tinyurl.com/y3bhurmm	https://tinyurl.com/rdkr4c7
Last download	6 December 2020	5 December 2020
Starting day	>1 January 2020	7 April 2020
End day	16 November 2020	6 December 2020
Type of publications	article, preprint, chapter, book monograph, preprint and proceedings	article, monograph, non-conventional and preprint
Fields and information provided	Bibliographic description Record provider Citations Altmetrics Open Access information	Bibliographic description Record provider
No. of records	168.053	118.200
No. of information sources	43	24
Main type and number of information sources	International Databases (2) Repositories (41)	International Databases (2) Repositories (10) Internal Databases (2) Others (10)
Main information sources and percentage of total records	Pubmed (47%) PMC (36%) medRxiv (4%) SSRN (4%)	Pubmed (51%) Internal database (30%) Elsevier (7%) medRxiv (6%)

Table 2. Contexts & scenarios: general view of the different timelines established

Dataset	**Time series name**	**Subseries an coverage periods)**	**Database**	**Forecast Starting and ending date**
General	TS1-General	TS1a - Total documents per day in WHO TS1b - Total documents per day in Dimensions	WHO Dimensions	07/11/2020 - 06/11/2021 14/10/2020 - 13/10/2021
Open Access	TS2- Access	TS2a - Total Open Access documents per day TS2b - Total Non-Open Access documents per day	“	“ “
Sources	TS3-Sources	TS3a - Total documents per day in Pubmed TS3b - Total documents per day in PMC TS3c - Total documents per day in meRxiv TS3d - Total documents per day in SSRN	“ “ “ “	“ “ “ “

The prediction of publication growth requires adequate tools to analyze historical data. There are several types of models that can be used for time-series forecasting. In this study we make use of ARIMA, which is a widely used forecasting method (Hyndman & Athanasopoulos, 2018). In the ARIMA model, only historical data of the variable of interest are used and forecasts are modelled as a linear combination of past observations and past error terms of the model (Hyndman & Khandakar, 2008). An ARIMA model is characterized by three parameters (p, d, q) where:

- p refers to the number of past values accounted in the model,
- d indicates the order of difference for attaining stationarity, and
- q specifies the number of error terms included in the model.

The ARIMA model can be used for non-stationary data, that is, for data in which the average and variance change over time. Since all eight time series exhibit a trend, the data are non-stationary and ARIMA handles non-stationarity by differencing subsequent observations. The necessary number of differencing to ensure stationarity is indicated by the parameter d. The three parameters are estimated from data, usually via maximum likelihood estimation (Hyndman & Athanasopoulos, 2018).

ARIMA models were fitted to the eight time series included in Table 2. All the analyses were conducted on an Ubuntu 18.04.1 machine, with R version 3.6.3 and RStudio version 1.1.456. The forecast analysis was carried out with a one-year window and specific results are offered for three-month windows. Along with point estimates, a 95% confidence interval accounts for the forecast uncertainty. Datasets and analyses of this study are openly accessible at https://doi.org/10.5281/zenodo.4478251.

Results

Evolution of COVID-19 scientific literature

The cumulated number of publications in Dimensions and WHO are presented in Figure 1. Dimensions indexed a total of 168,053 records and WHO a total of 118,200. As reported in Table 1, there are differences in the coverage of each source; while Dimensions covers records published in the last 10 months, WHO only does so for the last 8 months. Along with differences in size and period covered, we observe differences in the growth rate. In the case of Dimensions, it is more pronounced, especially from June onward. Both general time series fit a linear model, with R^2 values above 0.9 (R^2= 0.931 in WHO; R^2 = 0.851 in Dimensions).

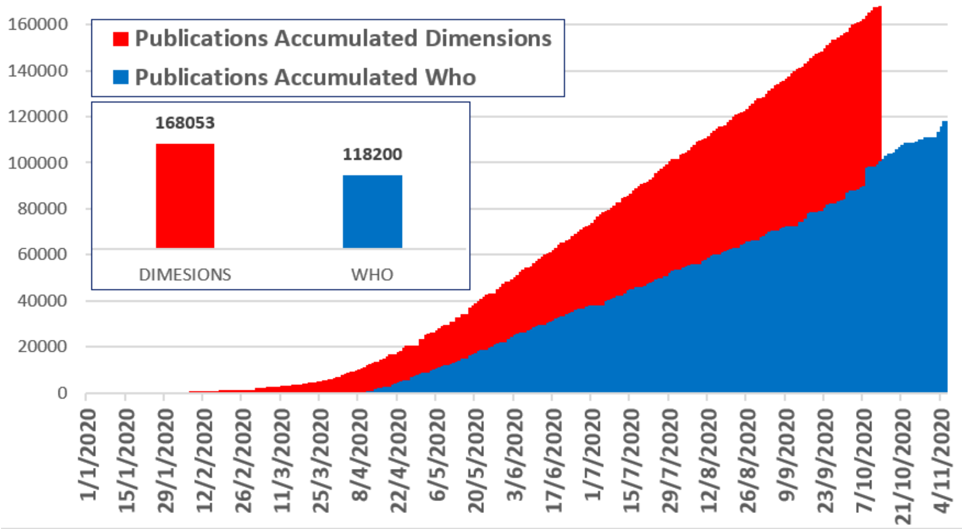


Figure 1. Accumulated number of records in Dimensions and WHO

Figure 2 shows the results for six time series. Figure 2A shows the results for Pubmed and PMC. These two repositories are the most prevalent sources in the Dimensions dataset, with PubMed alone including 47% of the share in this database (78,841 records). Figure 2B shows the time trend for medRxiv and SSRN. In this case we observe that both sources have similar volumes (7,002 and 6,002 records respectively) and a similar growth trend, with exponential growth until June 2020. Finally, Figure 2C compares the time series of OA and non-OA publications. Here the differences both in size and growth trends are very significant. OA literature is approximately five times larger than the non-OA and follows an exponential trend. In comparison, the growth of the non-OA publications is low.

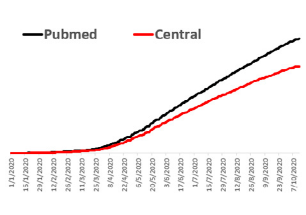
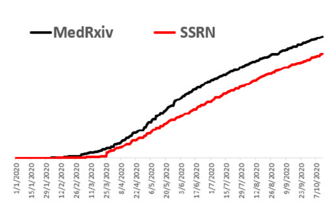
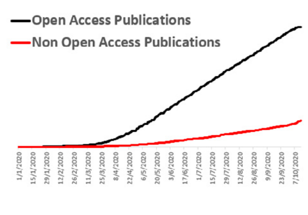
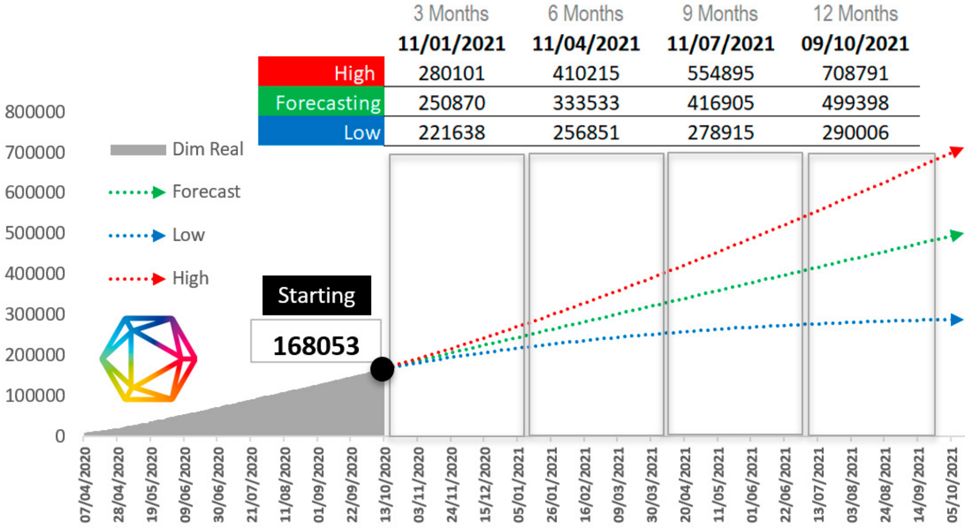
A Evolution in NLM products – PubMed and PMC	**B** Evolution in main repositories – MedRxiv and SRRN	**C**. Evolution by type of access – OA vs non-OA
		
PubMed: 78,841 PMC: 59,747	MedRxiv: 7,002 SSRN: 6,006	Open Access: 132,281 Non Open Access: 29,133

Figure 2. Time trend on the accumulated number of records in NLM databases, main repositories, and open access (OA)

Forecasting

Figure 3 and Figure 4 present the predictions for the Dimensions and WHO time series. We include our predictions along their uncertainty bounds. As observed, the lower bound shows a deceleration of growth, while in the two other cases it reflects a sustained rate of growth over time.

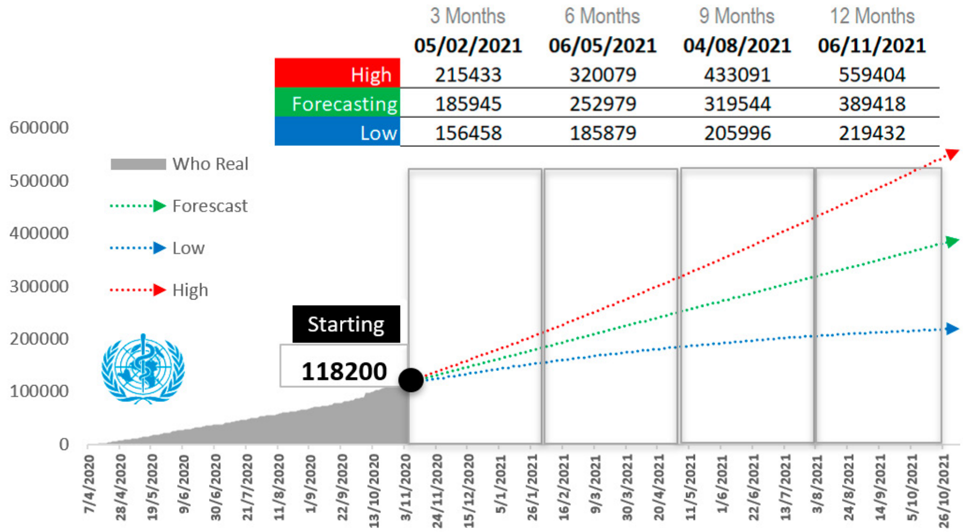


	3 Months 11/01/2021	6 Months 11/04/2021	9 Months 11/07/2021	12 Months 09/10/2021
High	280101	410215	554895	708791
Forecasting	250870	333533	416905	499398
Low	221638	256851	278915	290006

Figure 3. Forecasted growth of overall publications in Dimensions for 2021. Predicted growth (green) and upper (red) and lower bounds (blue) accounting for a 95% uncertainty interval. Forecasts are provided every three months

According to the ARIMA model, the forecast is that by the beginning of October 2021, the number of COVID-19 publications will reach half a million (499,398) according to Dimensions, with an upper bound of 708,791 records. This means that we expect the volume of COVID-19 publications to double by June 14th, 2021. If we consider the upper bound, the number of publications will double by February 20th, 2021.

A similar growth trend is observed for publications in the WHO database (Figure 4); the forecast is that 389,418 publications will be reached by the beginning of November 2021. The most likely maximum number of publications that is expected to be reached in the WHO database is 559,404. Based on the total number of records included on the date the data was collected, we should expect this number to double on June 11th, 2021. If we consider the upper bound of the forecast, the number of publications will double on February 24th, 2021 with 236,282. In both cases, the dates of growth and figures are similar, with Dimensions doubling the number of records in 7.8 months (243 days) and the WHO database in 7.13 months (217 days).



	3 Months 05/02/2021	6 Months 06/05/2021	9 Months 04/08/2021	12 Months 06/11/2021
High	215433	320079	433091	559404
Forecasting	185945	252979	319544	389418
Low	156458	185879	205996	219432

Figure 4. Forecasted growth of overall publications in the WHO for 2021. Predicted growth (green) and upper (red) and lower bounds (blue) accounting for a 95% uncertainty interval. Forecasts are provided every three months

Publication settings

Table 3 complements the general predictions in Dimensions and the WHO databases. The data is disaggregated and filtered based on three different settings: 1) type of access, 2) NLM source, and 3) domain-specific repository.

There are a total of 132,281 OA publications in the dataset (Table 3A). We observe an increase of 40% in their volume by the 14th of September, 2021. But the most intriguing growth is that of non-OA publications. Starting at an initial size of 29,133 at the time of the data retrieval, we expect an increase by a factor of 3.7 in the six following months, and 6.2 a year later. This spectacular increase is given by the rapid increase during the last period of registered data, as observed in Figure 2C. The upper growth scenario multiplies the starting non-OA papers by a factor of almost 11.

Similar forecast growth estimates are registered for PubMed and PMC (Table 3B). We estimate both sources will double their number of publications in a year. These two databases currently have a significant number of documents indexed, thus the effort required to double their size. Table 3C shows that these time windows are shorter for the two repositories analyzed, probably due to their smaller size. In the case of medRxiv, we estimate that the number of COVID-19 publications will increase by a factor of 15 in the next six months, and by a factor of 19 in a year (from 7,004 publications to 133,328). For SSRN, a more pronounced growth rate is estimated. In six months, the number of publications is expected to multiply by a factor of 17 and in twelve months by a factor of 25 (from 6,008 publications to 151,185).

Table 3. Forecast growth of publications by case scenario: A) type of access, B) NLM source and C) domain-specific repositories. It includes the predicted value and the upper bound of a 95% uncertainty level. Predictions are provided every three months

A *Time series by type of access (Open Access vs. non-Open Access)*					
Type	**Starting** 13/10/2020	**3 Months** 11/01/2021	**6 Months** 11/04/2021	**9 Months** 11/07/2021	**12 Months** 14/09/2021
OA	132,281	**155,661** High: 191,926	**176,705** High: 281,168	**197,983** High: 392,178	**219,027** High: 518,526
Non-OA	29,133	**81,482** High: 10,6783	**106,952** High: 151,899	**146,236** High: 228,054	**185,089** High: 309,963
B *Time series by NLM data source (PubMed vs. PMC)*					
Database	**Starting** 13/10/2020	**3 Months** 11/01/2021	**6 Months** 11/04/2021	**9 Months** 11/07/2021	**12 Months** 14/09/2021
PubMed	78.841	**98,879** High: 116,539	**118,236** High: 168,792	**137,808** High: 231,599	**158,025** High: 304,949
PMC	59.744	**74,644** High: 89,282	**89,321** High: 129,123	**104,162** High: 176,706	**119,492** High: 232,105
C *Time series by domain-specific repository (MedRxiv vs. SSRN)*					
Repository	**Starting** 13/10/2020	**3 Months** 11/01/2021	**6 Months** 11/04/2021	**9 Months** 11/07/2021	**12 Months** 14/09/2021
MedRxiv	7.004	**8,589** High: 10,849	**10,140** High: 16,618	**11,708** High: 23,735	**13,328** High: 32,174
SSRN	6.008	**8,259** High: 10,186	**10,525** High: 15,731	**12,817** High: 22,284	**15,185** High: 29,863

Discussion and concluding remarks

In this paper we present a forecasting analysis on the production of COVID-19-related scientific literature for 2021. We contribute to existing literature analysing the growth of science, a topic of interest since the very inception of scientometrics, with the pioneering works of Derek de Solla Price. However, we focus on a very particular type of scientific literature, that is, publications related to the COVID-19 pandemic. The scientific communication system has never generated as much interest, both scientific and societal, as it is generating during the COVID-19 crisis (Zastrow, 2020). Our results point towards potential scenarios for which infrastructure, communication strategies and policy actions must be coordinated to maximize the result of such paramount scientific effort (Brainard, 2020). We use the ARIMA model to predict literature growth as, despite the simplicity of this model, it proved to be highly accurate in our preliminary findings (Torres-Salinas et al., 2020). In times of social mistrust and fake news (Lazer et al., 2018), the production of new scientific knowledge must be accompanied by effective science communication strategies. The emergence of sources such as the WHO database and the CORD19 dataset already reflect a contribution to such efforts.

Although there is still debate as to what constitutes COVID-19-related literature (Kousha & Thelwall, 2020), the two databases have the unique feature of indicating daily indexing dates, which helps modelling data for predicting growth. Also, the level of transparency of these sources allows one to determine potential misrepresentations in certain fields (e.g., the inclusion of SocArxiv shows promise as to having a good coverage of social science fields). Our analysis by scenario points towards different levels (and, potentially, models) of growth depending on the data source used. Further steps will require looking into differences in growth rate by fields as well as considering external socio-economic and health factors which may affect the growth of scientific literature on this research front.

The urgency of the extraordinary health and financial crisis triggered by the pandemic has pushed the expansion of Open Acess and the inclusion of preprints as tacitly accepted scientific publications (although with many cautionary notes). This presents further challenges related to the control of scientific quality, certainty and rigour, although it is still too early to tell whether quality is being compromised in these pressing times of accelerated scientific discovery (Abritis, Marcus & Oransky, 2020). The fact that science is squarely in the social spotlight makes it especially vulnerable when errors are committed or when messages are misinterpreted. In the light of this framing, we believe that further research on this matter should continue to further our understanding of the growth not only of scientific publications, but also of the social reaction to science, and of the types of access by which scientific publications are made available.

Acknowledgments

This study is part of the project ‘Scientific communication in times of Corona virus’ funded by the TU Delft COVID-19 Response Fund.

References

Abritis, A., Marcus, A., & Oransky, I. (2021). An “alarming” and “exceptionally high” rate of COVID-19 retractions? *Accountability in Research*, *28*(1), 58–59. https://doi.org/10.1080/08989621.2020.1793675

Brainard, J. (2020). Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? *Science*. https://doi.org/10.1126/science.abc7839

Colavizza, G., Costas, R., Traag, V. A., Eck, N. J. van, Leeuwen, T. van, & Waltman, L. (2021). A scientometric overview of CORD-19. *PLOS ONE*, *16*(1), e0244839. https://doi.org/10.1371/journal.pone.0244839

Coles, J., & Eales, N. B. (1917). The history of comparative anatomy: A statistical analysis of scientific literature. *Science Progress*, *11*, 578–596.

Egghe, L., & Ravichandra Rao, I. K. (1992). Classification of growth models based on growth rates and its applications. *Scientometrics*, *25*(1), 5–46. https://doi.org/10.1007/BF02016845

Fernández-Cano, A., Torralbo, M., & Vallejo, M. (2004). Reconsidering Price’s model of scientific growth: An overview. *Scientometrics*, *61*(3), 301–321. https://doi.org/10.1023/B:SCIE.0000045112.11562.11

Homolak, J., Kodvanj, I., & Virag, D. (2020). Preliminary analysis of COVID-19 academic information patterns: A call for open science in the times of closed borders. *Scientometrics*, *124*(3), 2687–2701. https://doi.org/10.1007/s11192-020-03587-2

Hulme, E. W. (1923). *Statistical bibliography in relation to the growth of modern civilization*.

Kousha, K., & Thelwall, M. (2020). COVID-19 publications: Database coverage, citations, readers, tweets, news, Facebook walls, Reddit posts. *Quantitative Science Studies*, *1*(3), 1068–1091. https://doi.org/10.1162/qss_a_00066

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

National Science Board (2018). Science and Engineering Indicators 2018. NSB-2018-1. Alexandria, VA: National Science Foundation. https://www.nsf.gov/statistics/indicators/

Price, D. J. de S. (1963). *Little science, big science*. Columbia University Press New York. http://www.garfield.library.upenn.edu/lilscibi.html

Soltani, P., & Patini, R. (2020). Retracted COVID-19 articles: A side-effect of the hot race to publication. *Scientometrics*, *125*(1), 819–822. https://doi.org/10.1007/s11192-020-03661-9

Thelwall, M. (2020). Coronavirus research before 2020 is more relevant than ever, especially when interpreted for COVID-19. *Quantitative Science Studies*, 1–15. https://doi.org/10.1162/qss_a_00083

Torres-Salinas, D. (2020). Ritmo de crecimiento diario de la producción científica sobre Covid-19. Análisis en bases de datos y repositorios en acceso abierto. *Profesional de La Información*, *29*(2). https://doi.org/10.3145/epi.2020.mar.15

Torres-Salinas, D., Robinson-Garcia, N., & Castillo-Valdivieso, P. A. (2020). *Open Access and Altmetrics in the pandemic age: Forescast analysis on COVID-19 literature* [Preprint]. Scientific Communication and Education. https://doi.org/10.1101/2020.04.23.057307

Urbizagástegui, R., & Restrepo, C. (2015). The growth of Brazilian metrics literature. *Journal of Scientometric Research*, *4*(1), 1. https://doi.org/10.4103/2320-0057.156014

Zastrow, M. (2020). Open science takes on the coronavirus pandemic. *Nature*, *581*(7806), 109–110. https://doi.org/10/ggszkd

Zhou, P. (2013). The growth momentum of China in producing international scientific publications seems to have slowed down. *Information Processing & Management*, *49*(5), 1049–1051. https://doi.org/10.1016/j.ipm.2012.12.005

The causal intricacies of studying gender in science

Vincent Traag[1] and Ludo Waltman[2]

[1]*v.a.traag@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies (CWTS), Kolffpad 1, 2333BN, Leiden (The Netherlands)

[2]*waltmanlr@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies (CWTS), Kolffpad 1, 2333BN, Leiden (The Netherlands)

Abstract

The role of gender in science is a frequently debated subject. Papers on this subject use various terms to describe their findings, such as “gender bias”, “gender disparity”, “gender difference” or sometimes "gender gap". The different terms sometimes seem to be used interchangeably, making it unclear what researchers try to communicate with each term. To facilitate a clear discussion, we propose explicit definitions of these terms that emphasise the importance of the underlying causal mechanisms. Additionally, this causal language allows us to reason about concepts of fairness and their implications. The proposed terminology may contribute to a better understanding of the policy implications of a study.

Introduction

Science thrives on an open exchange of arguments and a plurality of perspectives. Scientific discussions should be open, frank and blind: only arguments should matter, not who presents them. Different viewpoints strengthen the scientific debate, and the inclusion of women and minorities in science will only contribute to this. Understanding the role of gender in science is crucial for improving the representation of women.

Gender differences in science are frequently studied in the literature. There are clear gender differences in citations, also when focusing on lead author positions (Sugimoto et al., 2013). A possible explanation of this might be gender differences in seniority: there are often more men than women in more senior positions. Other research corroborates this explanation and finds that gender differences in citations seem the result of gender differences in academic career trajectories and productivity (Huang et al., 2020). Another study attributed gender differences in citation to gender differences in journal prestige and collaboration patterns (Andersen et al., 2019). Similarly, gender differences in self-citation rates were reported (King et al., 2017), which are likely due to gender differences in careers and productivity (Mishra et al., 2018).

Women seem to transition into more senior positions less frequently, which can be partly explained by gender differences in productivity (Lerchenmueller & Sorenson, 2018). Although this is sometimes portrayed as a “leaky pipeline”, there seems to be a particular point in this pipeline in which these gender differences are most pronounced: the transition from postdoc to principal investigator (Lerchenmueller & Sorenson, 2018). After this transition, men and women seem to show similar career trajectories (Kaminski & Geisler, 2012; Hechtman et al., 2018). There is also evidence that men and women are not evaluated similarly when applying for academic positions, even when both have identical curricula vitae (Steinpreis et al., 1999).

Receiving funding is an important factor in making this transition towards principal investigator successfully. Experimental evidence suggests that gender identities on funding applications do not lead to gender differences in funding outcomes (Forscher et al., 2019). Other research suggests that gender differences in funding outcomes may depend on what criteria are used to evaluate the funding application (Witteman et al., 2019). An analysis of Dutch data suggests gender differences in funding rates (Van der Lee & Ellemers, 2015), although this may result from differences across fields (Albers, 2015).

Many findings point towards gender differences in productivity that may explain other observed gender differences. The gender difference in productivity is termed a “productivity puzzle” in earlier literature (Cole & Zuckerman, 1984). Some research suggests that articles submitted by women are reviewed differently than those written by men, and changing to double-blind review procedures attenuates gender differences (Budden et al., 2008). In contrast, recent research suggests that gender differences in publishing do not emerge as a result of being reviewed differently (Squazzoni et al., 2021). Although family formation, and related childcare, might be an explanation, early studies find no evidence of this (Cole & Zuckerman, 1987). There may be relevant field differences, where some fields have more intense periods around the time of family formation, while other fields may show such more intense periods at other times (Adamo, 2013). In math-intensive fields, family formation is suggested to be a key factor in explaining resulting gender differences (Ceci & Williams, 2011). Preliminary results from a large-scale survey suggests that women scientists indeed take on a caregiver role more frequently, although the implications on productivity are not clear (Derrick et al., 2019).

In our admittedly brief review of some of the literature on the role of gender in science, we intentionally used only the term “gender difference”. However, different studies use different terms to describe their findings. Several studies use the term “gender bias”, but its meaning is not always clear. Instead of “gender bias”, some studies use the term “gender disparity”, while others employ “gender difference” or occasionally "gender gap". The different terms sometimes seem to be used interchangeably, making it unclear what researchers try to communicate with each term. To facilitate a clear discussion, we propose explicit definitions of these terms. Defining such terms explicitly may help to clarify discussions about this topic.

As is clear from our brief review of some of the literature, many studies try to explain some gender differences by other gender differences. For example, gender differences in citation rates may be explained by differences in career trajectories and productivity (Huang et al., 2020). The idea of explaining gender differences by other gender differences is at the core of our definitions of gender bias, gender disparity and gender difference and builds on a fundamental concept: causality. We believe that the focus on causal reasoning may help to understand results in the literature, may sharpen the debate, and may lead to improved policy proposals.

The definitions of gender bias, gender disparity and gender difference that we propose are intentionally distinct from considerations of fairness. Nonetheless, this aspect is very important, and we also provide explicit definitions of fairness, distinguishing between procedural fairness and outcome fairness. Unlike our definitions of bias, disparity and difference, our fairness definitions rely on moral and normative judgements. The proposed definitions will help clarify some aspects of discussions around gender.

Structural causal models

The definitions we provide are based on the framework of structural causal models introduced by Judea Pearl (2009). Before we introduce our definitions of gender bias, gender disparity, gender difference and fairness, we need to introduce some of the concepts of structural causal models. We here only include a very brief introduction of the parts that are critical to our argument. For more information, please refer to Pearl (2009). A more accessible introduction is available from Dablander (2020) and a popular science account is provided by Pearl & MacKenzie (2018).

The basis for structural causal models is provided by directed acyclic graphs (DAG), which represent the causal relationships between variables. Each node in such a DAG represents a variable, and we may interchangeably use the term node or variable. We represent a link from a node X to a node Y by $X \rightarrow Y$, which represents a direct causal effect of X on Y. Acyclic means that there are no directed cycles, for example, if $X \rightarrow Y \rightarrow Z$ then a DAG may not contain the causal effect Z → X. Note that a DAG does allow X → Y → Z and X → Z, that is, there may be undirected cycles (i.e., ignoring the direction).

Any node X that has a directed path to Y is said to causally affect Y. Stated differently: if X were different, Y would also be different. The simplest example would be X → Y. Causality has a clear directionality: here X causally affects Y, but Y does not causally affect X. That is, if Y were different, this would not change X, since Y is a result of X, but X is not a result of Y. A simple example would be fire (X) and smoke (Y). Starting a fire will cause smoke, but creating smoke does not cause a fire. Note that there is a crucial difference between *observing* smoke and *creating* smoke: if we see smoke somewhere, it might be more likely that we can see a fire somewhere, whereas just creating smoke does not in itself start a fire.

The nodes that are directly affected by some node X are called its children, while the nodes that directly affect X are called its parents. For example, if X → Y we call X the parent of Y and Y the child of X. Similarly, children, children of children, and any nodes further downstream are called descendants. Parents, parents of parents, and any nodes further upstream are called ancestors. Hence, parents directly causally affect their children and ancestors causally affect descendants, possibly indirectly so.

Two variables may be associated, even if the one is not causally affected by the other. That is, two nodes X and Y may show some correlation or dependency, such that knowing X tells you something about Y and vice-versa. Determining which nodes are related is not straightforward, but DAGs are very helpful for this task. In order to properly explain how to determine whether two variables are associated, we need to introduce a number of other concepts.

d-connectedness and d-separation

The concepts of d-connectedness and d-separation indicate whether two variables are related or not in a DAG. The concepts of d-connectedness and d-separation may be somewhat difficult to comprehend. Below we explain these concepts briefly, but fully, because they are critical for some of our discussions around gender. For a more extensive introduction to d-connectedness and d-separation please see Pearl (2009), especially chapter 11.1.2 appropriately titled "d-separation without tears".

First of all, we require the concept of open or closed undirected paths between two nodes in a DAG. An undirected path consists of a sequence of nodes connected through links that may point in either direction. For example, X → Z → Y is an undirected path, but so are X ← Z → Y and X → Z ← Y. An undirected path is open if all nodes on the path are open. A node Z on a path is open if it is connected as ... → Z → …, called a mediator or if it is connected as ... ← Z → …, called a confounder, while a node Z is considered closed if it is connected as … → Z ← …., called a collider. Note that the same variable may play the role of mediator on one undirected path, a confounder on another undirected path, and a collider on yet another undirected path. Colliders will play an important role when we discuss the collider fallacy later in this paper. In short, undirected paths without any colliders are open while undirected paths with one or more colliders are closed. Instead of open or closed undirected paths, we will simply refer to these as open or closed paths.

In a sense, open paths allow information to flow freely between variables, while closed paths somehow block the information flow. If a pair of nodes X and Y is connected through at least one open path it is said to be d-connected, and information can flow freely between X and Y. If there are no open paths between X and Y they are d-separated, and no information can flow between X and Y. Two variables X and Y that are d-connected are associated. That is, if two variables X and Y are d-connected, knowing X tells you something about Y and vice-versa. Two variables X and Y that are d-separated are independent: knowing X tells you nothing about Y. The association between two variables that are d-connected does not need to reflect causality. The simplest example is X ← Z → Y, where the confounder Z affects both X and Y, so that X and Y are correlated only because of the common factor Z. In contrast, if X → Z → Y, the variable Z acts as a mediator and X and Y are d-connected, and the association between X and Y does reflect causality. Instead of open and closed paths others may sometimes use the terminology of unblocked and blocked paths.

In summary, variables are d-connected if there is at least one path with only confounders and mediators, see Figure 1 for an illustration. However, there is an important twist to d-connectedness and d-separation, which we discuss next.

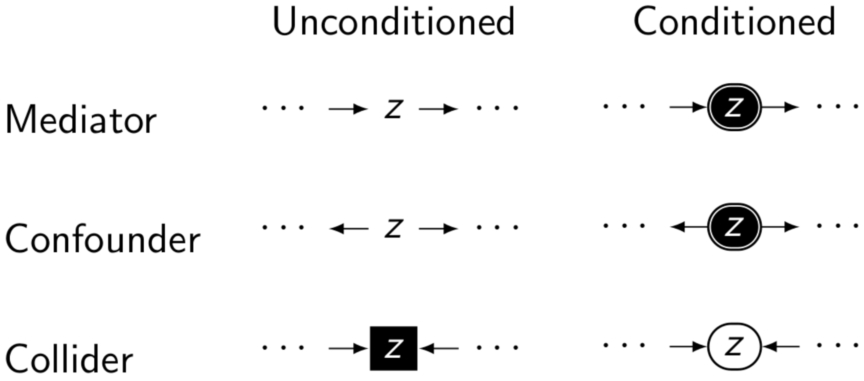


Figure 1. Illustration of when a node on an undirected path should be considered open (illustrated by white nodes) or closed (illustrated by black nodes). Conditioning on a variable "flips" a node from open to closed and vice-versa.

Conditioning and selection

Many studies condition on some variables. For example, quantitative studies frequently control for some variables by including them in a regression, which amounts to conditioning on those variables. Some studies may include only certain people in their analysis, for example considering only tenured professors or scholars that have published at least five publications. Such a type of selection amounts to conditioning on that variable. Other studies perform analyses on separate subsets of the data. A common example in science studies is for example analysing different scientific fields separately. Separating the analyses in distinct subsets amounts to conditioning on the variables that are used to define those subsets, such as the scientific field. Sometimes the fields are not analysed separately, but instead, some indicators, such as citations, may be field-normalised. Again, this amounts to conditioning on that variable. In short, conditioning on variables is a common sight, and it has profound implications for the notion of d-connectedness and d-separation.

When conditioning on a variable Z, a node that was considered open will become closed and vice-versa. That is, if a node Z on a path is conditioned[1] on, it is considered closed if it is connected as a mediator (… → **Z** →) or as a confounder (… ← **Z** → …) and considered open if it is connected as a collider (… → **Z** ← …). This is also illustrated in Figure 1. Hence, if a path is open, you can close it by conditioning on a mediator or confounder in that path, and the other way around, you can thus open a path that is closed by conditioning on a collider[2]. Because

[1] We will denote variables that are conditioned on in a path in bold.

[2] There is one additional complication. A collider becomes open not only if you condition on the collider itself, but also if you condition on any of its descendants. For example, if X → Z ← Y and also Z → **W**, then by

nodes may act as a mediator in one path and act as a collider in another path, it is possible that conditioning on a node closes one path yet opens another path.

Inequality, disparity and bias

We now have the necessary causal language in place to discuss the concepts of gender difference, gender disparity and gender bias in more detail. In particular, we suggest some definitions for these concepts, which we believe are useful for clearly conveying research results. These definitions are based on the framework of the structural causal models and directed acyclic graphs (DAG) that we just reviewed.

We propose to define a "gender difference" simply as any observed difference between people with a different gender. More formally speaking, suppose that the node G represents the variable gender in a DAG. Any node that is d-connected to G will then show a gender difference. Note that this may depend on whether nodes are conditioned on or not. Hence, some gender differences may appear when not conditioning on anything, while other gender differences appear only when condition on some variables.

Our proposal is to use the term "gender disparity" to refer to any difference between people with a different gender that is *causally affected by their gender*. This for instance means that if a woman had been a man (or vice-versa), the outcome of interest would have been different. In our causal language, if the node G again represents the variable gender, then any descendant Y of G (i.e. for which there exists a directed path $G \rightarrow X \rightarrow \ldots \rightarrow Y$) shows a gender disparity.

The strongest term is "gender bias", which we propose to define as any difference between people with a different gender that is *directly* causally affected by their gender. Similar to a gender disparity, this for instance means that if a woman had been a man, the outcome of interest would have been different. However, whereas a gender disparity may be the result of an indirect causal pathway from someone's gender to a particular outcome, a gender bias is a *direct* causal effect. In our causal framework, if G represents gender again, then Y shows a gender bias only if G is a parent of Y, that is $G \rightarrow Y$.

To clarify the distinction between a gender disparity and a gender bias, consider the example of being accepted at a prestigious university. Suppose that the acceptance rates for men and women are equal for each study programme, but that some study programmes have lower acceptance rates than others. Moreover, suppose there is a gender bias in study programme, and that women apply more often for study programmes with lower acceptance rates. This gender bias results in a lower overall acceptance rate for women. In this case, there is then a gender disparity in the overall acceptance rate. Because the causal effect is mediated by study choice, this gender disparity should *not* be called a gender bias in our terminology (see Figure 2). You may recognise this as an example of the famous Simpson's Paradox, which actually took place in Berkeley (Bickley et al., 1975). In contrast, suppose that a change in someone's gender on an application form affects the acceptance decision. In that case, gender does have a direct effect on acceptance, which means there is a gender bias in acceptance rates.

Note that the definitions of gender difference, gender disparity and gender bias are straightforward to apply to other variables of interest. For example, studies about the role of ethnicity, religion, or ideology may all use similar definitions of difference, disparity and bias.

conditioning on W we open the node Z. For more information about this, please refer to Pearl (2009), chapter 11.1.2.

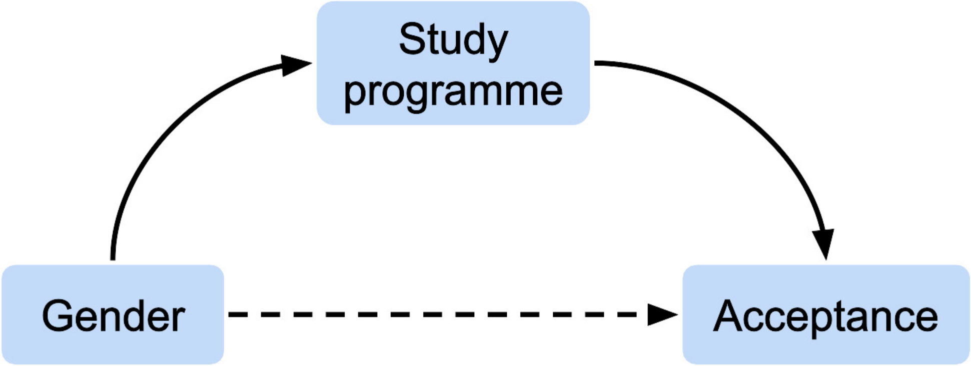


Figure 2. Example causal model. Gender causally affects the study programme, which causally affects acceptance. There is a gender bias in study choice, and a gender disparity in acceptance. If there is a direct causal effect of gender on acceptance (represented by the dashed line) there is a gender bias in acceptance.

Fairness

In colloquial language, disparities and biases are often seen as differences that are unfair. In the definitions we provided above, we intentionally did not include this moral or normative element. However, the issue of fairness is important and we try to clarify this notion here. A related debate is taking place in machine learning and artificial intelligence research, where fairness is an increasingly important consideration (Mitchell et al., 2021). Causal approaches similar to our definitions of gender differences, gender disparities and gender biases are also being developed (Zhang & Bareinboim, 2018; Makhlouf, 2020).

We believe it is pivotal to distinguish between two notions of fairness: procedural fairness and outcome fairness. Procedural fairness concerns the question whether a certain process is fair or not. In the example in the previous section, the acceptance decision depends only on the study programme: some programmes accept fewer applicants, and hence have a lower overall acceptance rate. In reality, such acceptance decisions are of course also influenced by other factors, such as aptitude or actual test scores. We then define procedural fairness as whether the direct causal effect of a variable on an outcome is seen as justified. If these direct causal effects are not seen as justified, the procedure for establishing the outcome is unfair. For example, most people would agree that using body height as a consideration in the acceptance decision for a study programme is unjustified, so this would be an unfair procedure.

An unfair procedure leads to an unfair outcome. However, the inverse is not necessarily true: The outcome of a procedure can be unfair, even if the procedure is fair. We define unfair outcomes as any outcome that results from an unfair procedure or that is causally affected by an earlier outcome that itself is unfair. Let us illustrate this in our previous example. Suppose that men and women have been swayed to apply to different study programmes, leading to the earlier identified gender bias in the choice of study programme. If we see this causal effect as unjustified, and hence as procedurally unfair, then the outcomes of the choice of study programme should be seen as unfair as well. This unfairness in the choice of study programme then propagates through the causal link and also renders the acceptance outcome unfair, even if the procedure used by a study programme to make acceptance decisions is fair.

More formally then, let the causal model again be represented by a DAG. Procedural fairness concerns the links of the DAG, and any direct link in the DAG can be labeled as unfair. Outcome fairness concerns the nodes of the DAG, i.e., the variables. Suppose that a direct causal effect $X \rightarrow Y$ is labeled procedurally unfair. Then the outcome Y is unfair with respect to X. Moreover, any descendant of Y also represents an outcome that is unfair with respect to

X.[3] In this definition, a single procedural unfairness X → Y may hence ripple through the network and render many outcomes unfair.

Procedural unfairness and outcome unfairness are related to gender bias and gender disparity. When there is a gender bias, gender has a direct causal effect on some outcome. If this causal effect is viewed as unjustified, it represents procedural unfairness, which results in an unfair outcome. When there is a gender disparity, but not a gender bias, gender has only an indirect causal effect on some outcome. The procedure for establishing that outcome may possibly be viewed as procedurally fair, but the outcome itself may still be unfair. Stated succinctly: Gender disparities may reflect outcome unfairness, while gender biases may reflect procedural unfairness. Whether something is indeed unfair remains a moral or normative decision. Different people may have different views on whether a specific gender disparity or gender bias reflects unfairness.

Interventions

The distinction between gender disparities and gender biases on the one hand, and outcome fairness and procedural fairness on the other hand helps to suggest where in the system an intervention may be more appropriate. To illustrate this, let us revisit the above example of being accepted at a prestigious university. If the effect of gender on acceptance rates is mediated by study choice, there is a gender bias in the choice of study programme, not in acceptance rates. If the effect of gender on the choice of study programme is viewed as unjustified, it would be procedurally unfair. An intervention targeted at study choice (e.g., making certain study programmes more attractive for women) addresses the procedural unfairness. By addressing the procedural unfairness, we also address any subsequent outcome unfairness that results from it, and address the root cause of the problem.

An intervention could also target the acceptance rates directly, for example by imposing a minimum acceptance rate for women or a certain quotum. Such an intervention does not address the original procedural unfairness, but seeks to rectify a previous wrong. It raises some thorny issues, with the pivotal question: do we intervene in a fair process in order to correct unfair outcomes? Doing so can be seen as an example of affirmative action or positive discrimination. The idea here is that the outcome of interest, the acceptance rate in this case, should be changed with respect to gender. In our causal model, this means that we would add a link G → A, where G and A represent gender and the acceptance decision. This additional link is intended to counter and nullify the undesirable gender disparity in the acceptance rate. Note that, in our terminology, there is now a gender bias in favor of women in the acceptance decision. This may raise other questions of fairness: if it is believed that gender should not have an effect on the acceptance decision, this additional link is procedurally unfair. In addition, introducing such a link may have some unexpected effects. The introduction of the link G → A leads to lower acceptance rates for men who choose the same study programmes as women. This might incentivise men to not choose programmes that traditionally attract more women. Instead of decreasing the gender bias in the choice of study programme, such an intervention may then actually reinforce this gender bias. In other words, even though the intervention may diminish the gender disparity in acceptance rates, it does not remedy the original procedural unfairness. There may still be good reasons for introducing such corrections of gender disparities, but their implications may be more complex than perhaps commonly thought.

Whether an intervention is desirable at all depends on whether the gender bias on study choice is seen as procedurally unfair. The concepts of outcome fairness and procedural fairness help to better understand the implications of different viewpoints that one may have on this issue.

[3] In fact, this line of reasoning can be generalized as follows. If X → Y is unfair, then Y or any descendant of Y is unfair with respect to any ancestor of Y. However, to simplify the discussion we do not consider this generalization here.

Whether a particular causal effect should be seen as procedurally unfair cannot be left to logic. It is a moral and normative judgment that should be subject to political and societal deliberation.

Collider fallacy

A gender difference does not necessarily reflect a gender disparity or a gender bias. Indeed, as we saw in the discussion on structural causal models, any two variables that are d-connected will show some association, but this does not need to represent a causal effect. The distinction between a gender difference on the one hand and a gender disparity and a gender bias on the other hand is inherently causal. Confusing a gender difference with a gender disparity or a gender bias complicates matters greatly in many studies. Collider fallacies may easily lead to such confusion.

To illustrate the problem of collider fallacy, we consider a recent paper about the role of gender in mentorship (AlShebli et al., 2020). The authors find that protégés with female mentors show a lower citation impact than protégés with male mentors. This paper was received quite critically and was recently retracted. Critics of the paper raised a number of concerns, for example about the data[4] and the operationalisation of the idea of mentorship (Lindquist et al., 2020). We illustrate the problem of collider fallacy on the basis of this study as a working example.

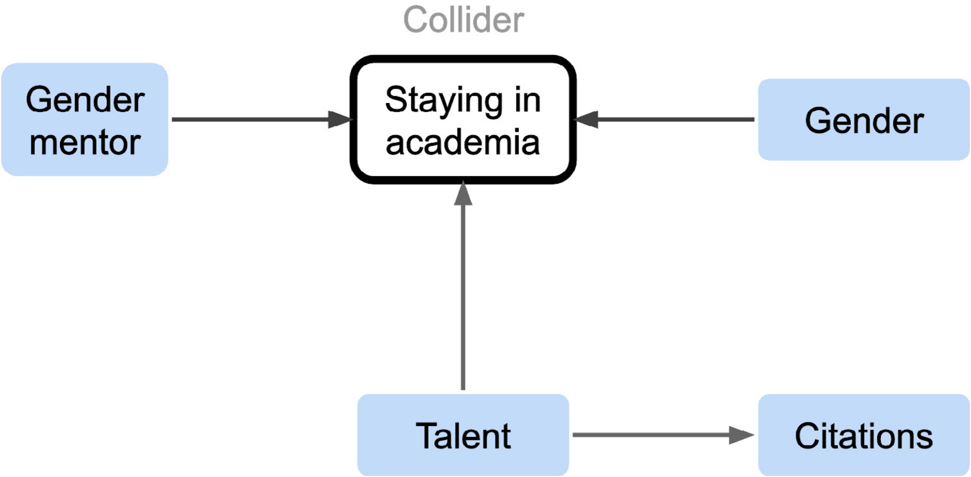


Figure 3. Simplified causal model of the role of gender in science. Each arrow represents a direct causal effect of one factor on another. For example, talent has a direct effect on staying in academia in this model.

We consider a simplistic causal model (see Figure 3) describing mechanisms relevant to interpreting the results by AlShebli et al. (2020). In our model, someone's research talent (T) affects both the citations (C) they receive and their likelihood of staying in academia (A). Independently of this, someone's gender (G) and the gender of their mentor (M) also affects their likelihood of staying in academia. More specifically, we assume that having a female rather than a male mentor makes it more likely for a female protégé to stay in academia.

In this causal model, staying in academia (A) is d-connected to citations (C) because of the path A ← T → C where talent acts as a mediator. This is the only path (longer than a single link) that is d-connected. All other paths are closed by node A, which acts as a collider for all other

[4] https://danieleweeks.github.io/Mentorship

paths. Hence, citations are independent from the gender and the gender of the mentor in this model. It could be discussed whether this is a realistic aspect of the model. However, our goal here is not to construct a fully realistic model, but rather to illustrate the concept of a collider fallacy.

AlShebli et al. (2020) make an explicit selection of the protégés included in the data collection: “we consider protégés who remain scientifically active after the completion of their mentorship period” (p. 2). As we explained before, this amounts to conditioning on the variable used to make a selection. If we condition on the variable “staying in academia”, this opens up a number of paths that were previously closed, leading to more pairs of d-connected nodes. For example, gender (G) will become associated with citations (C) because of the path G → **A** ← T → C. Moreover, the gender of the mentor (M) will become correlated with the citations (C) of the protégé because of the path M → **A** ← T → C. In other words, there will be a gender difference in citations, and also a difference in citations for the gender of the mentor. In our causal model, female protégés with male mentors are less likely to stay in academia, which means that those who do stay in academia can be expected to be more talented, on average, than their colleagues with female mentors. As a result, having female mentors is related to a lower research talent of protégés who stay in academia. Their lower research talent then in turn leads to fewer citations for those protégés. We would like to stress that this association does *not* reflect a causal effect. Instead, it is the result of conditioning on a collider, known as a collider fallacy. This example illustrates the problem of conditioning on colliders when studying causal effects. It leads to wrong conclusions. Depending on the extent to which our causal model captures the relevant causal mechanisms, the main result of AlShebli et al. (2020) may be due to this collider fallacy.

The possibility of a collider fallacy calls into question the policy recommendations made by AlShebli et al. (2020). The authors suggest that women should be paired with a male mentor because this has a positive effect on their citation impact. If the above causal model holds true, this suggestion is not correct. In this model, pairing a female protégé with a male mentor reduces the likelihood that the protégé stays in academia, which means that those protégés who do persevere in academia are likely to be more talented and to receive more citations. In our terminology: the difference between male and female mentors in the citations received by their protégés may be only a gender difference, not a gender disparity and certainly not a gender bias. Without additional evidence or assumptions, the observed gender difference does not support the policy recommendations made in the mentorship paper. In fact, given our conjectured causal model, it can be argued that one should do the opposite of what is suggested in the paper: to increase female participation in science, female protégés should be paired with female mentors. This illustrates the importance of considering the appropriate causal mechanisms for making policy recommendations.

Conclusion

Many papers on gender differences in science use a diffuse terminology and lack a clear causal framework. In an excellent and comprehensive review of the literature on gender differences in science funding, the lack of causal knowledge was identified as a sore point (Cruz-Castro & Sanz-Menéndez, 2020). The literature regularly discusses gender differences, disparities and biases without clear definitions. Moreover, without a clear causal model the findings may possibly lead to ill-conceived policy recommendations, which in some cases may actually hurt progress towards a better gender balance. We hope that our proposed definitions of gender difference, gender disparity and gender bias contribute to an improved appreciation of the causal intricacies in studying the role of gender in science.

Our proposed definitions of gender difference, gender disparity and gender bias do not include any moral or normative elements related to fairness. Instead, we define notions of fairness separately from the notions of difference, disparity and bias. We distinguish between procedural fairness and outcome fairness. Unfair procedures lead to unfair outcomes, which may render subsequent outcomes unfair. The proposed terminology enables us to reason in a systematic way about the consequences of unfair procedures, and whether to see outcomes as unfair. However, judging whether a procedure should be seen as unfair cannot be left to logic or algorithms, and should be the subject of a political and societal debate.

References

AlShebli, B., Makovi, K. & Rahwan, T. (2020). The association between early career informal mentorship in academic collaborations and junior author performance. *Nature Communications*, 11, 5855. https://doi.org/10.1038/s41467-020-19723-8

Adamo, S. A. (2013). Attrition of women in the biological sciences: Workload, motherhood, and other explanations revisited. *BioScience*, 63(1), 43–48. https://doi.org/10.1525/bio.2013.63.1.9

Andersen, J. P., Schneider, J. W., Jagsi, R. & Nielsen, M. W. (2019). Gender variations in citation distributions in medicine are very small and due to self-citation and journal prestige. *eLife*, 8. https://doi.org/10.7554/eLife.45374

Bickel, P. J., Hammel, E. A. & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175), 398–404. https://doi.org/10.1126/science.187.4175.398

Budden, A., Tregenza, T., Aarssen, L., Koricheva, J., Leimu, R. & Lortie, C. (2008). Double-blind review favours increased representation of female authors. *Trends in Ecology & Evolution*, 23(1), 4–6. https://doi.org/10.1016/j.tree.2007.07.008

Ceci, S. J. & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences of the United States of America*, 108(8), 3157–3162. https://doi.org/10.1073/pnas.1014871108

Cole, J. R. & Zuckerman, H. (1987). Marriage, motherhood and research performance in science. *Scientific American*, 256(2), 119–125. https://doi.org/10.1038/scientificamerican0287-119

Cole, J. R. & Zuckerman, H. (1984). The productivity puzzle : Persistence and change in patterns of publication of men and women scientists. *Advances in Motivation and Achievement*, 2, 217–258.

Cruz-Castro, L. & Sanz-Menéndez, L. (2020). *Grant allocation disparities from a gender perspective: Literature review. Synthesis report*. https://doi.org/10.20350/digitalCSIC/10548

Dablander, F. (2020). An introduction to causal inference. *BioRxiv*. https://doi.org/10.31234/osf.io/b3fkw

Derrick, G. E., Jaeger, A., Chen, P.-Y., Sugimoto, C. R., Van Leeuwen, T. & Larivière, V. (2019). Models of parenting and its effect on academic productivity: Preliminary results from an international survey. https://eprints.lancs.ac.uk/id/eprint/138455

Hechtman, L. A., Moore, N. P., Schulkey, C. E., Miklos, A. C., Calcagno, A. M., Aragon, R. & Greenberg, J. H. (2018). NIH funding longevity by gender. *Proceedings of the National Academy of Sciences of the United States of America*, 115(31), 7943–7948. https://doi.org/10.1073/pnas.1800615115

Huang, J., Gates, A. J., Sinatra, R. & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences of the United States of America*, 117(9), 4609–4616. https://doi.org/10.1073/pnas.1914221117

Forscher, P. S., Cox, W. T. L., Brauer, M. & Devine, P. G. (2019). Little race or gender bias in an experiment of initial review of NIH R01 grant proposals. *Nature Human Behaviour*, 3(3), 257–264. https://doi.org/10.1038/s41562-018-0517-y

Kaminski, D. & Geisler, C. (2012). Survival analysis of faculty retention in science and engineering by gender. *Science,* 335(6070), 864–866. https://doi.org/10.1126/science.1214844

King, M. M., Bergstrom, C. T., Correll, S. J., Jacquet, J. & West, J. D. (2017). Men set their own cites high: Gender and self-citation across fields and over time. *Socius: Sociological Research for a Dynamic World*, 3. https://doi.org/10.1177/2378023117738903

van der Lee, R. & Ellemers, N. (2015). Gender contributes to personal research funding success in The Netherlands. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(40), 12349–12353. https://doi.org/10.1073/pnas.1510159112

Lerchenmueller, M. J. & Sorenson, O. (2018). The gender gap in early career transitions in the life sciences. *Research Policy*, 47(6), 1007–1017. https://doi.org/10.1016/j.respol.2018.02.009

Lindquist, K., Gruber, J., Schleider, J. L., Beer, J., Bliss-Moreau, E. & Weinstock, L. (2020). *Flawed data and unjustified conclusions cannot elevate the status of women in science*. https://doi.org/10.31234/OSF.IO/QN3AE

Makhlouf, K., Zhioua, S. & Palamidessi, C. (2020). *Survey on causal-based machine learning fairness notions*. http://arxiv.org/abs/2010.09553

Mishra, S., Fegley, B. D., Diesner, J. & Torvik, V. I. (2018). Self-citation is the hallmark of productive authors, of any gender. *PLOS ONE*, 13(9), e0195773. https://doi.org/10.1371/journal.pone.0195773

Mitchell, S., Potash, E., Barocas, S., D'Amour, A. & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1), https://doi.org/10.1146/annurev-statistics-042720-125902

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (Second Edition). Cambridge University Press.

Pearl, J. & MacKenzie, D. (2018). *The Book of Why*. Basic Books.

Steinpreis, R. E., Anders, K. A. & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, 41(7–8), 509–528. https://doi.org/10.1023/A:1018839203698

Squazzoni, F., Bravo, G., Farjam, M., Marusic, A., Mehmani, B., Willis, M., Birukou, A., Dondio, P. & Grimaldo, F. (2021). Peer review and gender bias: A study on 145 scholarly journals. *Science Advances*, 7(2), eabd0299. https://doi.org/10.1126/sciadv.abd0299

Sugimoto, C. R., Larivière, V., Ni, C., Gingras, Y. & Cronin, B. (2013). Global gender disparities in science. *Nature*, 504, 211–213.

Witteman, H. O., Hendricks, M., Straus, S. & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *The Lancet*, 393(10171), 531–540. https://doi.org/10.1016/S0140-6736(18)32611-4

Zhang, J. & Bareinboim, E. (2018). Fairness in decision-making-the causal explanation formula. *AAAI Conference on Artificial Intelligence*.

The duality of disciplinary diversity for universities

Angelika Tsivinskaya[1]

[1] *atsivinskaya@eu.spb.ru*
European University at St. Petersburg, Center for Institutional Analysis of Science & Education, 6/1A
Gagarinskaya Street, 191187, St. Petersburg (Russia)

Abstract

In this paper, we investigate whether research and teaching mirror each other in terms of disciplinary structure. The diversity is presented in the conjunction of those two missions. Based on our assumption, we propose a typology that has two principal dimensions related to diversity in teaching and research respectively. As a measure of discrepancy between missions, we chose to assess the distribution of students and publications by scientific fields. We employed clustering analysis to establish disciplinary profiles of universities for publications and students separately. To test our typology we calculated overall diversity in teaching and research for each university and then applied clustering to identify possible types. Such an approach has the potential to reveal which disciplinary profiles have greater overall diversity. Our proposed methodology was applied to higher education institutions in Russia and we observed the greater diversity by scientific fields in publications comparing to students. Found empirical evidence on the connection between teaching and research can suggest that crossing disciplinary borders is easier in research than in teaching.

Introduction

Traditionally, education and research are viewed as being aligned in universities (Clark, 1995). In reality, the macrolevel missions are often coexisting, interlocking, or contradictory in nature (Scott, 2006). Moreover, we well know that the European higher education landscape is marked by a high level of heterogeneity (Huisman et al., 2015). So, it has raised the question of how Higher education institutions (HEIs) responded to increasing demands and being put under the pressure of not only teaching students but also conducting excellent research (Altbach, 2011). Therefore, we argue against the one-size-fits-all university model, which understands HEIs as isomorphic organizations dismissing the diversity of institutional types, and put the emphasis on the idea that universities should be evaluated with disciplinary specialization in mind (Bonaccorsi et al., 2021; Bornmann et al., 2013; Sánchez-Barrioluengo, 2014).

In this paper, we will assess the diversity both in research and in teaching by scientific fields for the whole universities populations on country level. We hope that our study will shed some light on the above issue by analyzing the patterns in research and teaching of Russian HEIs. To accomplish this goal, we use data from two different sources: our main source is the Russian Survey of Performance of Higher Education Institutions that includes data on each university, collected every year; additional data on the research activity are drawn by the RICS database, which is a national bibliometric database. We focus on the connection between diversity in research and teaching through an empirical study of Russian HEIs. Therefore, we present an empirical approach to the concept of diversity by addressing the following questions:

- Diversity of students: What is the distribution of students by educational programs?
- Diversity of publications: What is the distribution of published paper?
- Is there any connection?

Our paper contributes to the extant literature in several ways. First, we proposed a typology for HEIs, which focuses on the possible occurrence of mismatch between research and teaching missions. This gap is the starting point for our empirical analysis of diversity through those missions. The second contribution is related to reviewing the structure of the Russian higher education system using a combination of national databases that can give more comprehensive picture comparing to other sources especially in the case of establishing bibliometric profiles for universities.

Proposed typology

This typology arises from the debate about the alignment between research and teaching missions of universities. To explore the potential source of the discrepancy we chose to examine the distribution of students and publications by scientific fields. We argue that basic division for "generalist" and "specialist" represents only two "pure" types that have the agreement in two functions of research and teaching but in the higher education system can possibly exist universities of "mixed" types. Based on this we propose a typology that has two principal dimensions related to the degree of diversity in teaching and research by scientific fields. Theoretically, there are four possible types (Figure 1).

	Field-focused teaching	Diverse teaching
Field-focused research	type 2	type 3
Diverse research	type 4	type 1

Figure 1. Typology of universities.

Type 1 is what most people imagine talking about classical research universities, so to speak a "true generalist". We assume that most universities are in this category. However, are they representative of the whole higher education system in a particular country? In our case, we predict that big universities so-called flagships will be classified as type 1. These universities not only compete on the national level but also try to gain visibility on the international level.

Other "pure" type is type 2 which represents "true" specialist universities. As vestiges of the Soviet system are still present, some specialist universities continue to exist. They are the main suppliers of workers for domains of their respective ministries and as long as they get financial support from them and continue to be overlooked by separate ministries (Ministry of Agriculture, Medical Care, Transport) they stay in tacked and have more secure position than other universities.

Some would look down on the type 3 institutions as they teach different programs but have only a few that produce research. In many cases, these social programs with paid students considered a support system that gives money to subsidize programs that are able to produce research.

Institutions of type 4 are probably small ones but they make the effort to use external sources, creating projects with other organizations to spread their research but not able to establish they own teaching programs due to limited resources. Probably, they are oriented more towards research than teaching and do not have many students programs.

Data and methods

We have two main data sources are the Russian Science Citation Index (RSCI) which is a national citation database (Moskaleva et al., 2018) and the Russian Survey of Performance of Higher Education Institutions (Guba et al., 2020). The Russian Survey of Performance contains microdata available at the level of individual universities on a census basis about almost the entire population of HEIs in Russia. Initially, we had sample size=823 universities that have been registered as an organization in the RSCI. In this paper, we focused on bibliometric indicators such as the number of publications by scientific fields. Only scientific articles, reviews, short reports, conference proceedings and letters to the editor in journals were taken into account. As for the Russian Survey of Performance, it was used to obtain information about the number of students by study programs. According to the Survey, 769 universities were present in 2017, which was chosen as a target year for analysis. After cross-referencing two datasets and excluding universities with zero number of publications, the final version of the sample included 650 universities. Universities without publications are mostly small private universities oriented towards social sciences. In our final sample, around 74% are public universities. 560 universities are geographically located in regional capital cities. The median number of teaching and research staff is 224. As for median number of students, it is 3982.

For our analysis, we will use fields of science consisting of the following high-level groupings: Natural sciences, Engineering and Technology, Medical and Health Sciences, Agricultural Sciences, Social Sciences and Humanities. We transformed numbers of students and publications into percentages. We used the k-Means clustering algorithm for distributions of students and publications by scientific fields. We resorted to the pragmatic approach of using a sensible value for clusters K which is based on context information about the organization of the higher education system. In our specific case, a reasonable choice for K is a number of clusters equal to or at least comparable with that used in the Russian academic system which is the number K=6. To address our question about the connection between distributions of students and publications by fields, we calculated the Herfindahl-Hirschman Index (HHI) for educational programs and categories of papers. This measure denoted as is defined as:

$$\lambda(X) = \sum_{i=1}^{N} \left(\frac{x_i}{M} \right)^2 = \sum_{i=1}^{N} p_i^2 \text{ with } M = \sum_{i=1}^{N} x_i \text{ (the total number of items)}.$$

Findings

First, we show the result of cluster analysis for the distribution of students by scientific fields, than for publications. Second, we suggest types of profiles for universities which are based on HHI computed for students and publications and cluster assignment for universities employing these two measures as dimensions of differentiation. Finally, we present possible configurations of profiles of universities in the connection between the distribution of students and publications by scientific fields.

The following section shows the result of our clustering solution for distribution of students based on scientific fields. As you can see from Table 1, we have several clusters of universities with high specialization in taught programmes where more than 90% of all students are from one scientific field. These are clusters 1, 2 and 5, corresponding to oriented towards teaching in social sciences, medical sciences and humanities. Clusters 3, 4 and 6 are more diverse in terms of the body of students, but you still definitely prominent scientific fields which more than 50% of students are in them. Universities in the third cluster teach mostly students in the agricultural field but equally has around 20% of students in technical and social sciences. Universities in clusters 4 and 6 could be viewed as what could be called "classical" universities but it is apparent that 4 cluster has the base in social sciences and 6 cluster does in technical sciences. The first cluster is the biggest one and has 211 universities, whereas the second and the third are the smallest with around 50 universities in them. There is some percentage of students in social sciences in all clusters and the biggest number of universities is also the first cluster which oriented in social sciences.

Table 1. Cluster solution by students (cluster means, %).

Cluster	*N*	*fundamental*	*technical*	*medical*	*agricultural*	*social*	*humanities*
1	211	0.69	3.28	0.10	0.04	92.71	2.96
2	52	0.05	0.77	96.59	0.00	2.46	0.03
3	50	1.99	19.98	0.00	57.59	20.35	0.03
4	116	12.53	16.07	4.54	2.47	50.56	13.78
5	90	0.03	1.79	0.01	0.00	6.50	91.63
6	131	2.56	75.11	0.10	0.93	19.64	1.59
Between SS / total SS = 91.8 %							

In term of other characteristics, 75.6% of private universities in our dataset belong to cluster 1 which oriented towards social sciences. All medical universities are located in regional capital cities (cluster 2). The smallest universities with median numbers of students equal to 931 are in cluster 5.

Next, we present the cluster solution for universities but built upon the distribution of publication by scientific fields (Table 2). In comparison to cluster solution by students, it is more diverse in terms of the mixture of papers in different scientific fields, but you can still identify similar cluster considering dominating the scientific field for publication output. Cluster 2 and 6 can be two models for output of "classical" universities as was discussed above with cluster solution by students. Clusters 1, 3 and 4 are most field-focused, especially cluster 1 associated with medical sciences. In term of cluster sizes, we have three small ones – 1, 3, 5 and thee big ones – 2, 4, 6. The biggest one is cluster 6 with publications mostly in social sciences.

Table 2. Cluster solution by publications (cluster means, %).

Cluster	*N*	*fundamental*	*technical*	*medical*	*agricultural*	*social*	*humanities*
1	65	6.01	0.82	79.90	0.59	8.88	3.05
2	129	23.45	40.71	2.54	2.33	23.59	5.60
3	63	0.83	1.30	0.99	0.09	20.44	74.19
4	166	1.52	3.27	1.31	0.89	85.79	5.67
5	56	7.58	8.27	2.58	53.51	22.13	3.67
6	171	11.04	6.62	5.41	2.77	49.11	21.93
Between SS / total SS = 83.2 %							

Our next step is to view the possible configuration of universities based on overall diversity in two dimensions without specificity of considering scientific fields which we presented earlier. We empirically test whether these configurations present in our higher education system. For this purpose, we calculated the Herfindahl-Hirschman Index (HHI) for distributions by educational programs and by papers and viewed the validity of those configurations (Figure 2).

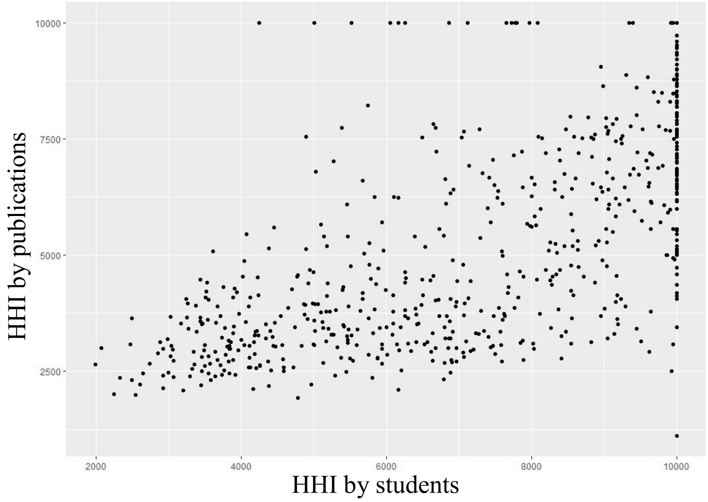


Figure 2. Diversity of disciplinary profiles by HHI.

Using the results of the cluster analysis HEIs can be classified into three clusters. So, we can conclude that universities which have students in different fields but produce papers only in one are rare. On the most basic level, the diversity of taught programs could be assumed as a supportive environment that generates diversity in research too.

Finally, we would like to show whether those profiles of universities based on different approaches have great consolidations among them or only universities with a specific body of students and publications have an inherent propensity to a specific configuration. Using flow diagram we present connections between different cluster solutions (Figure 3). You clearly can see that cluster assignments for universities based on the dominating scientific field by students and by publications are in close concordance for universities oriented towards agricultural, medical, technical-classical sciences and less for social, humanities and social-classical, so to speak it is the "great divide" between natural and social sciences.

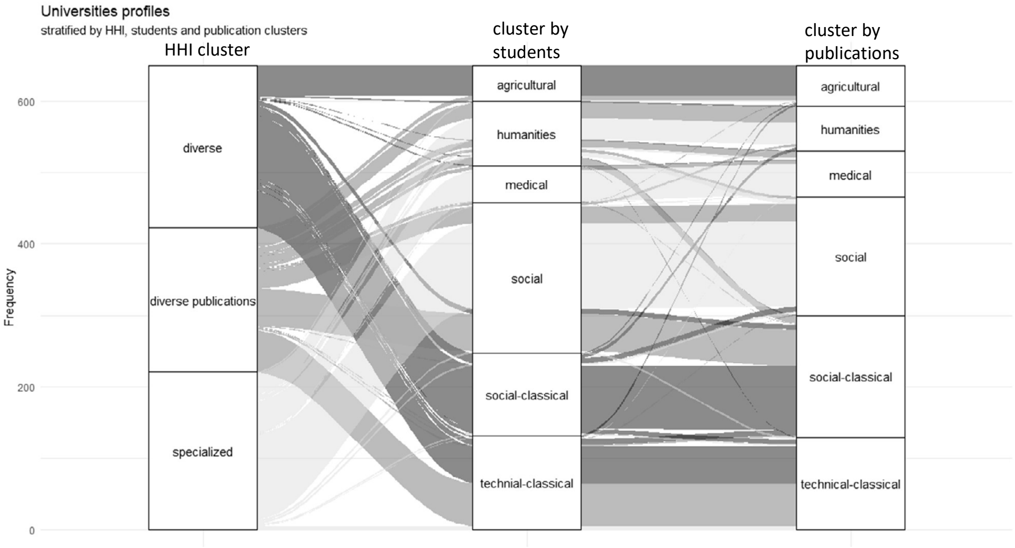


Figure 3. Connections between disciplinary profiles.

To sum up the overview of the insights we gained from our analysis are:

- The high number of universities with the social field as the main profile with specialization in it.
- The universities with different study programmes and publications concentrated only in one field are rare. External links can be detected (for example, technical universities publishing a paper in medicine).
- Migration to publish in close scientific fields (technical universities generate research in fundamental sciences, researchers in social sciences and humanities are also easily cross the border between scientific fields).

Conclusion

The greater diversity in publication comparing to teaching programs can have several possible explanations. Firstly, it is easier to publish a paper in the other field than to establish a teaching program at university. There is no restriction or evaluation beforehand for publishing paper in a different field. For example, using some available data on medical condition, the researcher in applied mathematics can write a relevant paper for a medical journal. Secondly, the greater diversity in publication can be a result of direct collaboration between institutions or with industry. In this case, the data can be supplied by one organization when other is more involved in analyzing it. Thirdly, the classification for teaching programs is more restricted than for publications; however, we aggregated our data on the quite high level in term of scientific fields where this effect should be partially mitigated.

This study can be used to recognize for instance week and strong scientific fields and this information can enrich the knowledge about HEIs in Russia. The findings of this research can be used in several ways by HEI administrators: for redistribution of existing funding system of universities, for restructuring universities teaching programs and for identification of universities' mission. Our findings are not only informative on the individual institutional level but as well as on the HEI system on the whole.

Our analysis evaluates the research performance of universities considering the multi-disciplinary nature of institutions and the sustainability of teaching-research nexus. It is innovative in taking this perspective on higher education in Russia.

This study has some limitations. Our findings are based only on Russian universities so they cannot be generalized on the other HEIs but the suggested research approach can be implemented in other cases with some revision of suitable data sources. In this paper, we did not measure or discuss the quality of publications or educational programs. Another important topic which is not covered but we have great interest to explore in the future research is the temporal aspect of diversity.

References

Altbach, P. G. (2011). The past, present, and future of the research university. *Economic and Political weekly*, 65–73.

Bonaccorsi, A., Belingheri, P., & Secondi, L. (2021). The research productivity of universities. A multilevel and multidisciplinary analysis on European institutions. *Journal of Informetrics, 15*, 101–129.

Bornmann, L., Moya Anegón, F. de, & Mutz, R. (2013). Do universities or research institutions with a specific subject profile have an advantage or a disadvantage in institutional rankings? *Journal of the American Society for Information Science and Technology, 64*, 2310–2316.

Clark, B. R. (1995). *Places of inquiry: Research and advanced education in modern universities*. Univ of California Press.

Guba, K., Sokolov, M., & Tsivinskaya, A. (2020). Fictitious Efficiency: What the Russian Survey of Performance of Higher Education Institutions Actually Assessed. *Voprosy obrazovaniya / Educational Studies Moscow*, 97–125.

Huisman, J., Lepori, B., Seeber, M., Frølich, N., & Scordato, L. (2015). Measuring institutional diversity across higher education systems. *Research Evaluation, 24*, 369–379.

Moskaleva, O., Pislyakov, V., Sterligov, I., Akoev, M., & Shabanova, S. (2018). Russian Index of Science Citation: Overview and review. *Scientometrics, 116*, 449–462.

Sánchez-Barrioluengo, M. (2014). Articulating the 'three-missions' in Spanish universities. *Research Policy, 43*, 1760–1773.

Scott, J. C. (2006). The Mission of the University: Medieval to Postmodern Transformations. *The Journal of Higher Education, 77*, 1–39.

Writing style and success in research grant applications

Peter van den Besselaar[1,a] and Charlie Mom[2]

[1] *p.a.a.vanden.besselaar@vu.nl*
Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam (Netherlands)

[2] *charlie@teresamom.com*
TMC research, Middenweg 203, 1098 AN Amsterdam (Netherlands)

Abstract

Why are some research grant applications selected for funding and others not? In this paper we investigate whether the writing style influences the evaluation scores and the decision of selection panels. A linguistic analysis of the grant proposal abstract, the project description and the CV of the applicant do reveal several characteristics of the texts that have a positive effect on the score the grant applications receive, and through that on the probability to be selected for funding. The study suggests that writing style does have moderate a effect on application success, and therefore should not be neglected by applicants.

Introduction

Why are some grant proposals selected, and others not? From a normative perspective, it should be based on merit (Merton 1973), and evaluation panels are expected to base their decisions on the performance of the applicants, and on the quality, novelty, and originality of the proposed research project. However, peer review does not work in this way (Chubin & Hackett 1990; Lamont 2009, Thorngate et al. 2009) and various forms of bias may enter decision making, based on e.g. gender, or on social or cognitive links between the applicant and members of the selection panel.

This paper focuses on other non-merit characteristics of a grant application that may play a role in the evaluation of grant applications: The textual quality of the application. Is it well written, comprehensible, and to the point? And does it show that the applicant has the capabilities to do the proposed research? One may of course argue that peers when evaluating grant proposals should and can look through the presentational form and are able to evaluate the scientific content, but there are indications that this is not always the case. For example, in several grant application procedures, applicants have to present their project in front of the selection panel, and research shows that the scores before and after that presentation can be very different (Van den Besselaar & Van Arensbergen 2013). As applicants and universities are increasingly aware of this, applicants for prestigious career grants are often trained by communication professionals to give a strong and convincing presentation for the panel. If the 'stage performance' influences the evaluation of grant applications, this may also be the case for 'textual performance'. In this paper we focus on the *application text* and answer the question which textual characteristics have effects on the evaluation of the proposal. Does the way a grant proposal is written influence the fate of the application?

Only recently, the impact of language use in science has become a topic of research within science studies. Some research has focused on language use in *scientific articles*. In a recent paper, Lu et al. (2019a) studied the relationship between scientific writing and scientific impact, using linguistic complexity indicators as a proxy for the quality of scientific writing. The study is based on 36400 full text Biology articles and 1797 full-text Psychology articles and found (practically) no significant relationship between linguistic complexity and citation impact. In another paper, Lu et al. (2019b) compared the writing style of native authors with non-native authors in terms of (i) syntactic complexity, such as sentence length and sentence complexity, and (ii) lexical complexity, such as lexical diversity, density, and sophistication. Only marginal differences were found between non-native and native English speaking authors. Hengel (2020) also analyzed the writing of manuscripts, but with another aim. Comparing the submitted manuscripts with the accepted paper, she showed that (i) women submit better written manuscripts than men do, but still (ii) women improve the manuscript during the revision period even more than men do. This suggests that women are held to higher standards in the peer review process than men are. An indirect effect could be that women have to invest more time per paper, which in turn could partly explain the differences in scientific output between male and female authors (Hengel 2020).

Where Hengel used article manuscripts to study gender bias in peer review, Kaatz et al. (2014, 2015, also Magua et al. 2017) did so using linguistic analysis of *peer review reports*. They showed systematic differences in the words used in the review reports about female applicants compared to those about male applicants, suggesting the possibility of gender bias. Van den Besselaar et al. (2018) analyzed the language of *review reports* with the aim of identifying which evaluation criteria are deployed and how the selection is done. One finding is that negative word categories (negation terms and negative evaluation terms) are much stronger predictors of the score applicants receive than superlatives and positive evaluation words, suggesting that panels are much more focused on eliminating applications that are considered weak than on finding the excellent and promising proposals (see also Hume et al. 2015). Another finding was the negative correlation between the percentage of project and research related words, and the received score. This is in line with the theory of Festinger (1950) that disagreement among panel members leads to more discussion in the panel and to lower scores.

Another recent strand of research focuses on the language use in *grant applications*. Do writing style and the textual quality of the application play a role? Explorative qualitative work was done to identify the characteristics of grant writing as a genre (Connor & Mauranen 1999; Feng & Shi 2004). In a recent small-scale study, Boyack et al. (2018) found a positive relation between writing quality of grant applications and application success. In a follow-up study, the same authors used five linguistic variables to predict grant success: (i) proposal size, (ii) the *Gunning fog index* for writing clarity, (iii) male-oriented language, (iv) positive emotional language, and (v) the *story arc* of emotional words. Out of these, only the *story arc* had a significant positive effect on grant application success (Smith et al. 2019). The authors distinguished between six types of *emotional story lines* and found that these result in different average success rates. In a study of anonymous grant applications to the Gates foundation, Kolev et al. (2020) found that even when the sex of the applicant is hidden, women score systematically lower than men, also after controlling for e.g., the topic of research. This suggests that the writing style differs between men and women, and that the way of writing has an effect on the evaluation of the applications. Indeed, the study found that women use more specific words which correlates negative with review scores and success rates. Men use more general words, and those correlate positive with scores and success. Obviously, reviewers like specific project descriptions less than more general claims, and men profit from that.

Markowitz (2019) studied the relation between the writing style and the size (in UD$) of the grant, using a linguistic approach. The study is based on the idea that language reflects psychological characteristics and cognitive abilities (e.g., Tausczik & Pennebaker 2010). The writing style communicates characteristics of the author, such as whether he/she is able to communicate clearly (comprehensibility), able to write and think about complex issues (intellectual capacities), and whether he/she has sufficient self-confidence (for doing the work promised in the application), and these dimensions are expected to be important when evaluating grant applications or conference submissions (Markowitz 2019, 2021).[b]

Based on these considerations, Markowitz (2019) formulated the hypothesis that reviewers may be more convinced by proposals that show (i) a writing style that reflects complex and analytic thinking by the applicant, (ii) clear writing, and (iii) (self-)confidence in the proposal. Text complexity was measured by the number of words (longer texts are more complex), by the number of words per sentence (longer sentences are more complex) and by the use of analytic language. Clarity of the text is measured through the share of common (English) words in the text: the higher that share, the less technical terms (jargon). Finally, confidence in the proposal is measured through the number of certainty terms (pointing at more self-confidence), tentative terms (pointing at a lack of self-confidence) and causal terms (causal terms reflect stronger claims in the proposal). Using a dataset spanning multiple decades and containing approximately 20000 granted proposals across several directorates of the NSF, these linguistic characteristics were expected to predict the grant size.

Most of the hypotheses of Markowitz were supported, but some were not: Common words and analytic language do not positively but negatively affect the size of the grant. These findings may suggest an alternative interpretation of what counts as complex writing and thinking, and what counts as easy to read. Easy to read text is probably not text with more common words, but text with a good narrative structure (Jones & McBeth 2010), and a higher score for narrative writing implies a lower score on analytical writing (Pennebaker et al. 2014). A negative score on common words does not imply that the text is more difficult to read, but that it contains more technical terms, adding to the complexity of the text. Therefore, our expectations would go in the opposite direction: Firstly, a grant application with a more narrative style (= a lower score on analytical writing) receives a higher evaluation score and has a higher probability to get funded, and secondly, a grant application with more technical terms (a lower score on common words) receives a higher evaluation score and has a higher probability to get funded.

The hypothesis that tentative words have a negative effect on the grant size was supported by Markowitz' analysis, but we find that unexpected because tentative words have different functions in scientific texts. Tentative words may indeed indicate that a researcher is uncertain, hesitant, and indecisive, which may negatively influence the assessment of a grant proposal. On the other hand, it is also good practice to use tentative words indicating that findings are always provisional and may be falsified in the future, and that conclusions should be formulated cautiously (Hyland 1996). And tentative words used in the latter way may be positively received by reviewers. These two roles of tentative terms may balance out. We would therefore expect that an application text with more certainty terms and more causal words receives a higher evaluation score and has a higher probability to get funded, but that tentative terms do not have such an effect.

In this paper, we address several limitations of the Markowitz study. (1) We also include *rejected applications*, as the comparison of rejected and successful applications may better show the effect of writing style on grant decisions, than only relating writing style of awarded grants with the size of the grant. (2) We use not only the *abstracts* of the applications, but also other important parts: the *project description*, and the *CV*. The three texts have different functions and may therefore have different writing styles (Feng & Shi 2004). Especially CVs have a distinct structure, and often consist of listings such as previous academic positions, grants, and publications. However, in our case the CV also includes a page where the applicants summarize their main achievements in a narrative. Based on this, we expect that the effects of the linguistic variables will stronger in the project proposal than in the short abstract, and in the abstract stronger than in the CV. (3) All applications are written in English, but most applicants are not native English. One would expect that the quality of the texts also depends on *English language proficiency* of the author, which is included as control variable. (4) We link the analysis to the decision-making process. Our case consists of a two-step procedure, and the impact of language may be different in the two phases. One may expect that the writing style is more important in the first round when a large number of applications have to be evaluated in a short time, than in the second round when a smaller number of only the best are left for the final selection. A panel member may then go in depth through the smaller number of remaining applications – even when these are sometimes not written too well. More generally, this suggests that what counts as adequate text is *context dependent*. We would therefore expect that language does have a much smaller effect in the second step of the procedure than in the first step.

Case description

The dataset consists of 3207 applications for an early career grant, covering the year 2014. The applications are handled by 25 disciplinary panels, covering all disciplines. The overall success rate is 11.7%, indicating that the funding scheme is very competitive. The maximum grant is 1.5 million euro, for 5 years. Almost everyone received the same amount of funding. The process of getting the grant consists of two steps. In the first step proposal and CV are reviewed, and a score is given for both the PI and the proposal. This is translated into a C, B, or A score. Only the proposals with and A proceed to the second step, where another round of reviewing (by more reviewers) takes place and the PI is interviewed by a panel. The final score in the second step can be B (not fundable) of A (fundable), and a part of the proposals that received an A are funded.

Data and methods

The applications cover all fields, grouped into 25 panels. The dataset is almost complete, as 95% of applicants gave informed consent for this study. We deleted one panel for which most of the project descriptions were lacking. After removal of five unreadable proposals documents, the dataset consists of 2900 application texts. The applications are about 11 pages long, with (i) an abstract of one page, (ii) a project description of five pages and (iii) a curriculum vitae (CV) of five pages. The PDF files were converted into TXT, and then split into the three parts. We checked whether the file splitting was done correctly. For example, all file-parts that were small or big in relation to the expected length were manually checked. Where needed, the split was redone manually. In quite some cases this was indeed needed. We also found some other data problems. In some other cases, we manually corrected the data: this was done where a wrong structure of the text misled the automatic parsing of the application texts, and for the application texts that included an empty front page, or included a table of contents. Furthermore, quite some abstract texts that still included formatting instructions were cleaned, as the instruction text would change the linguistic structure of the abstract considerably.

The Abstracts, Project descriptions, and CVs were analyzed with the Linguistic Inquiry and Word Count tool (LIWC), using the 2015 English dictionary (Seih et al. 2017). We deploy the same LIWC categories as described above for operationalization of *discourse and thinking style complexity* (1-3), *clarity of the text* (4), and of *confidence in the proposal* (5-7), but some with another interpretation then in earlier work (Pennebaker et al. 2014; Markowitz 2019):

1. Word count: The longer the text, the more information provided. However, too long texts may be counterproductive, and therefore we also use *word count squared*. The relation between the length of the text and the score may be an inverted U-curve.
2. Words per sentence: Longer sentences reflect more complex arguments. And also here, too long sentences are making the text unnecessary complex and may have a negative effect on readability.
3. Analytic thinking: A higher score on this composite linguistic indicator reflects stronger *complex and analytic thinking*, while a lower score reflects stronger *storytelling and narrative thinking*.[c]
4. Common words: The LIWC dictionary contains common words, but not technical terms – so the higher the score, the less technical terms in the text. One may interpret this as an indicator of readability of a text, but also as in indicator of complexity: the higher the score on common words, the smaller the technical content of a text, and the lower its complexity.

5. Certainty terms, expressing the level of confidence the applicant has in the proposal.
6. Tentative terms, which are interpreted in the Markowitz study as an expression of *hesitation*. However, tentativeness may also mean that the text is *prudent* in its arguments and conclusions. We would expect that these two meanings of tentative terms work in opposite direction, and therefore not relate to the panel score.
7. Causal language indicate that the proposed project has strong claims (Connor & Wagner 1998) and therefore may be more convincing for the panel members or reviewers.

The case covers one year only, so possible changes in language use over time do not affect the analysis. The same holds for possible changes in the implicit or explicit criteria that are used to evaluate the grant proposals. As this study is restricted to a single funding instrument, the dataset is rather homogeneous. This may help to isolate the effect of language use more precisely. In order to reflect possible disciplinary differences in writing, the linguistic variables were standardized at panel level. Furthermore, a few covariates are included. (1) Panels is used as random effect in the mixed model analysis, and as covariate in the ordinal and the logistic regressions. (2) The applicants come from many countries, and only a small minority has English as mother language. Therefore, English language proficiency may therefore differ considerably between the applicants. Using the reports of a large international language training institute (Education First 2013), we created a score for English proficiency at the country level. This is a rather rough measure, so one should interpret the findings with care.

As dependent variables we use the score for the project proposal, the final score, and success of the application. To predict the scores for the *project proposal*, we conduct a multi-level analysis with SPSS26, using linear mixed models, and panel as random effect. The *final score* is an ordinal variable with five values, from low to high: C, B (first step), B (second step) A, A-funded, which were recoded it into numerical values 1 to 5. We use ordinal regression to test whether the linguistic variables affect this score, with panel as covariate to control for disciplinary differences. Finally, we use the dichotomous variables *to-step2/not-to-step2* and *grant/no-grant* as dependent variables, and for predicting these, we use logistic regression. The analyses are done for step 1 and for step 2 of the procedure, as the role of language may differ between the two phases in the procedure. Each of these analyses is done for each of the three texts (project, abstract, and CV).

Results

As the project text is much longer than the abstract, we use the former one for the core analysis. The analysis is repeated for the abstract and for the CV, which may strengthen the findings. Table 1 shows for the project text the correlation between the main variables. The score received in Step-1 for the project proposal correlates positively with word count, certainty terms, causal terms, and the level of English language proficiency. It correlates negative with words per sentence, common words and analytic thinking. There is no correlation with tentative terms. This confirms most of the expected relations.

Table 2 shows the results of the multivariate (mixed models) analysis how the linguistic variables affect the project score in the 1st step of the section procedure, where we control for the level of English language proficiency and include panel as random effect. The analysis shows that more complex texts (longer text, longer sentences and less common words) work positively, but too long text and too long sentences work negatively. And a more narrative style works positively on the score, just as the use of certainty words and causal words, and as expected, tentative words do not play a role. English language proficiency has also a positive effect on the project score.

Table 1: Pearson correlations between the main variables (project text, step 1)

1	Project score	1	2	3	4	5	6	7	8	9	10	11
2	CV score	.838**										
3	Word count	.309**	.288**									
4	WC^2	.274**	.257**	.980**								
5	Words/sentence	-.097**	-.079**	-.058**	-.047*							
6	W/S^2	-.110**	-.091**	-.067**	-.053**	.984**						
7	Common words	-.064**	-0.035	-.068**	-.081**	.434**	.403**					
8	Analytic thinking	-.119**	-.169**	-.062**	-.048*	.142**	.143**	-.060**				
9	Certainty words	.057**	.071**	-0.011	-0.022	.137**	.129**	.265**	-.114**			
10	Tentative words	0.026	0.03	0.015	0.007	.037*	0.034	.275**	-.290**	.216**		
11	Causal words	.053**	.046*	-0.023	-.040*	.039*	0.027	.227**	-.148**	-0.01	.098**	
12	English proficiency	.154**	.176**	.039*	0.027	-.056**	-.058**	0.027	-.237**	-0.003	0.012	.072**

** Correlation is significant at the 0.01 level (2-tailed). * Correlation is significant at the 0.05 level (2-tailed). N=2900

Table 2. Project score by linguistic characteristics of project text, step 1

	Estimate	Std. Error	Df	t	Sig.	95%CI-lb	95%CI-ub
Intercept	2.5954	0.0238	23.6	108.95	0	2.5461	2.6446
Word count	0.5025	0.0478	2875.8	10.50	0	0.4087	0.5963
WC^2	-0.3461	0.0478	2875.8	-7.24	0	-0.4398	-0.2524
Words/sentence	0.1391	0.0535	2875.7	2.60	0.009	0.0343	0.2439
W/S^2	-0.1630	0.0527	2875.7	-3.09	0.002	-0.2662	-0.0597
Common words	-0.0452	0.0114	2875.8	-3.95	0	-0.0676	-0.0227
Analytic thinking	-0.0249	0.0103	2877.7	-2.41	0.016	-0.0451	-0.0047
Certainty words	0.0406	0.0099	2875.8	4.11	0	0.0212	0.0600
Tentative words	0.0026	0.0103	2875.9	0.26	0.796	-0.0175	0.0228
Causal words	0.0283	0.0098	2875.8	2.90	0.004	0.0092	0.0474
Engl. proficiency	0.0686	0.0098	2898.7	7.00	0	0.0493	0.0878

Mixed models, with panel as random effect. All independent variables were standardized at the panel level.

Table 3. Final score by linguistic characteristics of project text

	Estimate	Std. Error	Wald	df	Sig.	95%CI-lb	95%CI-ub
[Score = 1]	-0.306	0.193	2.522	1	0.112	-0.684	0.072
[Score = 2]	1.763	0.196	81.154	1	0	1.379	2.146
[Score = 3]	2.191	0.197	123.462	1	0	1.805	2.578
[Score = 4]	2.805	0.201	195.714	1	0	2.412	3.198
Word count	1.583	0.221	51.163	1	0	1.149	2.017
WC^2	-1.073	0.212	25.543	1	0	-1.489	-0.657
Words/sentence	0.451	0.208	4.705	1	0.03	0.044	0.859
W/S^2	-0.51	0.208	6.048	1	0.014	-0.917	-0.104
Common words	-0.153	0.043	12.667	1	0	-0.237	-0.069
Analytic thinking	-0.148	0.038	14.902	1	0	-0.223	-0.073
Certainty words	0.155	0.037	17.541	1	0	0.083	0.228
Tentative words	-0.018	0.039	0.209	1	0.648	-0.093	0.058
Causal words	0.065	0.037	3.149	1	0.076	-0.007	0.137
English proficiency	0.224	0.037	37.083	1	0	0.152	0.296

Ordinal regression. Nagelkerke Pseudo R2 = 0.146;
Panel scores not included in the table.

The proposals also get a *final score* which classifies them into five groups. Do the linguistic variables affect the finale score? Table 3 shows the results of the analysis, which are very similar to the results of the mixed models analysis reported above: Complex texts which are longer, have longer sentences, more technical terms (= fewer common words), and use a narrative style (a lower score on analytical language) get higher scores. Causal language and certainty also have a positive effect, as has English proficiency. The use of tentative terms has again no effect on the score. The explained variance is 0.146 (Nagelkerke pseudo R^2).

The linguistic variables may also influence the *decisions*. In the first step, the decision is whether a proposal is selected to progress to the second step. About 25% of all proposals go into the second step, where half of them receive the grant. A logistic regression shows that most of the independent variables have similar regression coefficients as in the previous analyses (Table with results not included in this paper). The explained variance (Nagelkerke pseudo R^2) is 0.106. The findings of the various analyses support our first three expectations.

Above, the expectation was formulated that the effects of the linguistic variables would be much smaller in the second step of the procedure than in the first step, and the findings support this. In step 2, most of the correlations between the project score and the linguistic variables are not significant (table 4). In the mixed model analysis, the length of the proposal has a positive effect on the evaluation score, but too long texts score lower. And the share of common words has a negative effect on the project evaluation score. All the other linguistic variables and English language proficiency do have no effect. Finally, only the word count variables have a significant effect on the decision in step 2.

Table 4: Pearson correlations between the main variables in step 2 (project text)

1	Project score (step 2)	1	2	3	4	5	6	7	8	9	10	11
2	CV score (step 2)	.691**										
3	Word count	-0.018	-0.02									
4	WC^2	-0.027	-0.026	.990**								
5	Words/sentence	-0.028	-0.029	-0.066	-.076*							
6	W/S^2	-0.02	-0.024	-0.066	-.072*	.985**						
7	Common words	-0.048	-0.016	-.107**	-.131**	.382**	.355**					
8	Analytic thinking	0.004	-.104**	0.024	0.029	.099**	.091*	-.072*				
9	Certainty words	0.017	0.065	-0.069	-.083*	.090*	.079*	.274**	-.109**			
10	Tentative words	0.007	0.051	-0.019	-0.028	0.066	.072*	.299**	-.293**	.239**		
11	Causal words	-0.016	-0.011	-.098**	-.099**	0.044	0.038	.181**	-.160**	-0.002	0.035	
12	English proficiency	0.021	0.028	-0.018	-0.022	-0.021	-0.022	-0.006	-.184**	0.001	-0.011	0.013

** Correlation is significant at the 0.01 level (2-tailed). * Correlation is significant at the 0.05 level (2-tailed). N=754

The previous analyses were done on the *project text*. A similar analysis could be done for the *abstract* text and the *CV* text. Does the writing style of the applicant in these texts have the same effect on scores and decisions? We lack space to show these analyses, but the patterns are similar, although less pronounced.

Further research

Despite the fact that current CVs are rather structured with many listed items and not so much normal text, similar relations between word use and evaluation scores and decisions were found as for the project descriptions. However, recently one observes a trend to ask for a more discursive CV instead of merely lists of positions, experience, and (bibliometric) achievements (VSNU 2019). The linguistic properties of successful CVs are a topic for further research.

This paper studied how elements of self-presentation affects the evaluation scores and application success. However, different social groups may present themselves in different ways, which may affect the evaluation outcomes differently. In a future study we plan to extend the analysis by including more characteristics of applicants, especially gender. We also plan to include other elements of self-presentation such as gendered language use.

Conclusions

In contrast to earlier studies, this study includes *also the non-successful applicants*, and is based on *different texts* belonging to grant applications (abstract, project, CV). The findings suggest that writing style has an effect: The way of writing influences the scores grant applications receive, and through that the probability of success. The linguistic characteristics of the texts do have an effect on the panel scores, on the final score, and on the grant decision. Complex language (longer text and longer sentences), with technical content (lower share of common words) is beneficial, and positively affects the scores and the probability to get funded. In contrast to earlier work (Pennebaker et al. 2014), a more *narrative writing style with technical terms* works better. Finally, confidence in the grant proposal expressed through *certainty* terms and *causal* language has a positive effect, but in contrast to earlier work (Markowitz 2019) we do not find any effect of tentative terms. This was expected, as tentative terms can have two functions: expressing uncertainty but also expressing the necessary tentative nature of scientific findings.

These findings are most visible when analyzing the project text, but the analysis of the proposal abstract and of the CV point in the same direction – this despite the fact that these texts have other functions (Feng & Shi 2004) and, especially the CV, also a different structure.

The study enabled to distinguish between two steps in the project selection process, and language and writing style have an impact in the first phase of the selection process, but not anymore in the second. In the first phase, the writing style and readability may be important as many applications have to be processed by the selection panel within a relative short period, whereas in the second phase more time is available to assess the remaining smaller set of applications. Above that, the applications that make it to the second phase are already perceived by the selection panel as excellent, and therefore the reviewers may be more inclined to carefully read the text even if the writing style is not too good. This suggest that what counts as an adequate text is context dependent.

The linguistic model to predict the final score (using ordinal regression) explained a moderate part of the variance (Nagelkerke pseudo $R^2 = 0.146$). And using logistic regression to predict the first selection decision (proceed to Step 2 or not) resulted in a Nagelkerke pseudo R^2 of 0.1. Concluding, the effect of the writing style may not be very strong, but it is substantial.

Finally, this study suggests that the role of the writing style may be context dependent. We showed that for the two phases of the application procedure, but it may also be the case for e.g., phases in the career: what is perceived as a good text of a novice may be different from what is perceived as a good text of an expert.

Acknowledgments

This paper is a result of the GendERC project, which received funding from the ERC under grant 610706. Contract partners were Joanneum Research (Vienna, Austria), Vrije Universiteit Amsterdam (Amsterdam, the Netherlands), and Tecnalia (Bilbao, Spain). The authors thank Anne Baker (University of Amsterdam) for support on how to interpret the linguistics results, and David Markowitz (University of Oregon) for comments on an earlier version of the paper.

References

Boyack, K. W., Smith, C., & Klavans, R. (2018). Toward predicting research proposal success. *Scientometrics*, 114, 449–461.

Chubin D., E., & Hackett E. J., (1990). *Peerless science: peer review and U.S. science policy*. State University of New York Press.

Connor U & Mauranen A (1999) Linguistic Analysis of Grant Proposals: European Union Research Grants. *English for Specific Purposes,* 18, 47–62,

Connor, U., & Wagner, L. (1998). Language use in grant proposals by nonprofits: Spanish and English. In: *New Directions for Philanthropic Fundraising*, 1998, 59–74.

Education First (2013). *EF English Proficiency Index.* https://www.ef.edu/epi/

Festinger, L. (1950). Informal social communication. *Psychological Review,* 57(5), 271–282

Feng, H., & Shi, L. (2004). Genre analysis of research grant proposals. *LSP and Professional Communication,* 4, 8-32.

Flanagin AJ (2007) Commercial markets as communication markets: uncertainty reduction through mediated information exchange in online auctions. *New Media & Society,* 9(3), 401–423.

Hengel E. (2017). *Publishing while female. Are women held to higher standards? Evidence from peer review*. Cambridge Working Papers on Economics 1753.

Hume KM, Giladi AM & Chung KC (2015) Factors impacting successfully competing for research funding: an analysis of applications submitted to the Plastic Surgery Foundation. *Plast Reconstr Surg,* 135(2), 429e-435e.

Hyland K (1996). Writing Without Conviction? Hedging in science research articles. *Applied Linguistics,* 17, 433-454

Jones MD & McBethand MK (2010), A narrative policy framework. Clear enough to be wrong? *The Policy Studies Journal,* 38, 329-353.

Kaatz A, Gutierrez B & Carnes M (2014). Threats to objectivity in peer review: The case of gender. *Trends in Pharmacological Sciences,* 35(8), 371–373.

Kaatz A, Magua W & Zimmerman DR, Carnes M (2015). A Quantitative Linguistic Analysis of National Institutes of Health R01 Application Critiques From Investigators at One Institution: *Academic Medicine,* 90(1), 69–75.

Kolev J, Fuentes-Medel Y& Murray F (2020). *Is blinded review enough? How gendered outcomes arise even under anonymous evaluation*. NBER

Lamont M (2009). *How professors think. Inside the curious world of academic judgement*. Harvard University Press.

Larrimore L, Jiang L, Larrimore J, Markowitz DM & Gorski S (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research,* 39, 19-37

Lu C, Bu Y, Dong X, Wang J, Ding Y, Larivière V, Sugimoto CR, Paul L &Zhang C (2019a). Analyzing linguistic complexity and scientific impact. *Journal of Informetrics,* 13, 817–829

Lu C, Bu Y, Wang J, Ding Y, Torvik V, Schnaars M & Zhang C (2019b). Examining Scientific Writing Styles From the Perspective of Linguistic Complexity. *Journal of the Association for Information Science and Technology,* 70, 462–475.

Madera GM, Hebl MR & Martin RC (2009) Gender and letters of recommendation for academia: agentic and communal differences, *Journal of Applied Psychology,* 94, 1591–1599

Magua W, Zhu X, Bhattacharya A et al. (2017) Are Female Applicants Disadvantaged in National Institutes of Health Peer Review? Combining Algorithmic Text Mining and Qualitative Methods to Detect Evaluative Differences in R01 Reviewers' Critiques. *Journal of Women's Health,* 26, Nr 4.

Markowitz DM (2019). What Words Are Worth: National Science Foundation Grant Abstracts Indicate Award Funding. *Journal of Language and Social Psychology*, *38*(3), 264–282.

Markowitz, DM (2021). Words to Submit by: Language Patterns Indicate Conference Acceptance for the International Communication Association. *Journal of Language and Social Psychology,* 49 – online first

Merton RK (1973) The normative structure of science. In: *The sociology of science*, U Chicago Press, 267-278

Pennebaker JW, Chung CK, Frazee J & Lavergne GM (2014) When small words foretell academic success: The case of college admissions success. *Plos One,* 9(12), e115844

Pennebaker (2015). Reply to a comment on Pennebaker et al. 2014.

Seih YT, Beier S & Pennebaker JW (2017). Development and Examination of the Linguistic Category Model in a Computerized Text Analysis Method. *Journal of Language and Social Psychology*, 36(3), 343–355.

Smith C, Boyack KW & Klavans R (2019). Towards predicting proposal success, an update. *Proceedings ISSI Conference.* Roma, 770-781

Tausczik & Pennebaker JW (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology,* 29, 343–355.

Thorngate W, Dawes RM & Foddy M (2009) *Judging merit.* New York: Psychology Press.

Van den Besselaar P & Van Arensbergen P (2013). Talent selection and funding of research. *Higher Education Policy,* 26, 421-427.

Van den Besselaar P, Sandström U & Schiffbaenker H (2018), Using linguistic analysis of peer review reports to study panel processes. *Scientometrics,* 117, 313-329

VSNU (2019). Room for everyone's talent. Towards a new balance in the recognition and rewards of academics (a publication of VSNU, NFU, KNAW, NWO and ZonMw). https://vsnu.nl/files/documenten/Domeinen/Onderzoek/Position%20paper%20Room%20for%20everyone%E2%80%99s%20talent.pdf (Accessed on 25/12/2020)

[a] Currently scientific advisor at DZHW, Schützenstraße 6a, 10117 Berlin (Germany)

[b] Examples outside the science domain where linguistic characteristics of text are relevant for the decision are e.g., applications for loans (Larrimore et al. 2011) or for jobs (Madera et al. 2009).

[c] This composite indicator is almost the same as the CDI (Pennebaker et al. 2014): "LIWC2015's analytical thinking is based on the CDI. The LIWC variable has been normalized so that the mean is 50 and scores range from 0 to 100. However, CDI and analytical thinking should correlate .99 with each other" (Pennebaker, reply November 25, 2015).

What leads to gender bias in review panels?

Peter van den Besselaar[1, a] and Charlie Mom[2]

[1] *p.a.a.vanden.besselaar@vu.nl*
Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam (Netherlands)

[2] *charlie@teresamom.com*
TMC Research, Middenweg 203, 1098 AN Amsterdam (Netherlands)

Abstract

Gender bias in selection panels for positions and grants has been studied intensively over the last decades. However, apart from comparing review scores received by men and women, we do not know much about how panel composition, procedures and processes influence selection. This is an omission, as decision making on grants and academic positions is strongly influenced by the prioritizing and selecting in those panels. Based on the relevant social-psychological literature on small group decision making, we distinguish several panel attributes that may influence the decisions in the panel in general and the prevalence of gender bias more specifically: (i) Composition of panels in terms of gender composition and variety of countries represented in the panel, (ii) workload of the panels, (iii) and especially the level of gender stereotyping in the panels. This explorative case study suggest that panel composition indeed correlates with gender bias: the more women the stronger the gender bias against women, the more experience male panelists have, and the more countries represented in the panel, the stronger the bias in favor of women. Workload did not correlate with gender bias, but the level of gender stereotyping does: the stronger stereotyping as found in the review reports, the stronger bias against women applicants.

Introduction

Gender differences within science may have become smaller over time, they have not disappeared. Women are still underrepresented in higher academic positions (Ginter & Kahn 2006; Kahn & Ginter 2018), apply less often for grants (Dickson 1997) and have – not undisputed - lower probability to receive grants (Holman et al. 2018). Many of the relevant decisions with respect to positions and grants are made in panels. Panel members are expected to deploy a variety of merit related selection criteria (Van Arensbergen et al. 2014). However, as social-psychological research on group decision making has shown, also a variety of small group (panel) characteristics does influence the decision outcomes (Olbrecht & Bornmann 2010; Van Arenbergen et al. 2014b), and bias seems unavoidable. Research on review panels is largely qualitative (Lamont 2009), but in this paper we use a quantitative approach to detect what panel characteristics predict gender bias in panels. A core mechanism that we include is in this paper is *gender stereotyping*, that is panelists implicitly assuming men are better suited for science than women (Ellemers 2018; Madera et al. 2009; Sekaquaptewa et al. 2003: Leslie et al. 2015), and which leads to differences in how men and women are assessed. The first aim of this paper is to determine what panel characteristics lead to gender bias. The second aim is to measure gender stereotyping in order to find out whether the level of gender stereotyping predicts the level of gender bias at the panel level. We do this using a case study on gender bias in grant allocation.

Research on bias in grant allocation

Gender bias in grant peer review was put on the research agenda and the research policy agenda by the Wennerås & Wold (1997) study of the Swedish medical research council, a paper which has become a point of reference in the debate. However, the literature on gender bias in grant allocation has resulted in contradicting findings, partly suggesting the absence of gender bias in grant allocation (Ceci et al. 2014; Williams & Ceci 2011, Friessen 1998; Grant et al. 1997;

Dickson 1997; Hosek 2005; Sandstrom & Hallsten 2008; Jayasinghe et al. 2003; Marsch et al. 2008, 2009, 2011; Beck & Halloin 2017), and partly suggesting that gender bias is a real issue in grant decisions (Wennerås & Wold 1997; Bornmann et al. 2007; Ban der Lee & Ellemers 2015; Van den Besselaar et al. 2018; Witteman et al. 2019). However, two problems exist with most of these studies, and these problems may also explain the contradictory results. Firstly, many studies do not include merit (or performance) variables, which is generally acknowledged as a major drawback. As Marsh et al. (2008) argue, "An important limitation in our research program with grant applications, and in peer-review research more generally, is that we had no fully appropriate external criteria against which to validate the outcomes of the peer-review process. We argued that the final panel decision (based on the integration of all available information) was the best criterion available for validating and testing potential biases in the ratings by individual assessors. However, a more external criterion is needed to validate the results of the panel decision itself. (…….) However, at least in terms of ratings of researcher quality, the previous track record of researchers does provide a viable validity criterion. Nevertheless, the need to find more suitable validity criteria remains a critical issue for research into the review of grant applications and for peer-review research more generally"
The other drawback is that the analysis is generally done at the level of a council or a funding instrument, without taking the panel level into account, despite the fact that at the panel level the decisions are made or at least heavily influenced. In this paper we specifically focus on panel decision making and on the panel characteristics that influence the level of gender bias. And more particularly, we focus on measuring the effect of (unconscious) gender stereotyping on gender bias.

Research question

In this paper we address the following research question: What panel characteristics explain the differences in gender bias between the panels? The question is important for a better understanding of the dynamics of - within the science system widely used - panel decision making. Better understanding of these dynamics may also contribute to improving the quality of decision making.

The case

Our case is the 2014 Starting Grant of the European Research Council (ERC). The sample consists of 3030 (out of 3200) applicants are distributed over 25 disciplinary panels. The panel processes in this case are not formalized, neither are the criteria formulated for the panelists. According to the council, the only criterion that should be deployed is the level of excellence of the project and the principle investigator. This is described in the following way: Excellent investigators have shown creative independence thinking, the ability to do ground-breaking research, and have moved beyond the state of the art. As panels members are excellent researchers in their respective fields, the panelists should therefore decide among themselves what excellent applications are. But interviews with panel members have shown that they do not find it easy to operationalize the concept of excellence and its elements such as 'independence' or 'the ability to do groundbreaking research'. This results in uncertainty and in different approaches by different panelists. Panelists doubt about criteria deployed and express the need for clearer and operational criteria for 'excellence'. Inspecting the review reports support this uncertainty. In some cases, a PI will get a low score because his/her "publications are not well cited", in others because "publications are not in the main multidisciplinary journals", and again in others because "the proposal has a high risk", despite that the ERC asks for "high risk – high gain" projects. Additionally, the workload of the panels is rather high, which may lead to heuristics-based decision making, and these heuristics may be stereotype-based (Van Arensbergen et al. 2014b).

The case is interesting, as it contains of panels in which women have a lower success rate than men, panels with an about equal success rate, and panels where women have a higher success rate then men. However, success rates as such are an indicator of gender differences, but not necessarily of gender bias. In order to measure bias, one needs to control for differences in merit – which we did elsewhere for 22[b] out of 25 panels using logistic regression (Van den Besselaar & Mom - forthcoming). The dependent variable was the panel decision, and several independent merit variables were included: (i) Performance variables measuring productivity, impact, and earlier grants), reputation variables such as publications in top journals, ranking of the host institution, and the median ranking of former employers of the applicant, and organizations the applicant collaborated with; and personal characteristics such as age, academic age, and of course gender. Several other relevant merit-variables cannot be included here due to a lack of data: e.g., *prizes and awards*, and *community roles* such as support roles in the work environment, editorships of journals, board member of learned organization, or program chair of conferences.

In order to inspect the gender effect in more detail, the predicted margins (or predicted probabilities: PP) were calculated using STATA 16. When using logistic regression, this is needed as it provides the effect of gender for individual values of independent variables. This would remain unclear from the regression coefficient alone – ass the latter gives the gender effect only for one value (zero) of the independent variables (Mood 2009).

For each panel, a *gender bias indicator* was calculated: the average predicted probability of women (PPw) in a panel divided by the success rate of women in that panel (SRw). When this indicator is larger than one, there is bias against women. When the indicator is smaller than one, the panel shows bias in favor of women. And the higher the value of the indicator, the stronger the bias against women. The advantage of the indicator is that it allows for bias in both directions, as there may be panels with bias against women and panels with bias in favor of women. Indeed, using the gender bias indicator defined above, we do find bias *in different directions*: six of the panels show bias against women, seven show bias in favor of women, and nine other panels are about neutral. Table 1 shows the results of the analysis at panel level.

Table 1. Locus and effects of gender bias at panel level (N=22).

Panels	*All*	*LS*	*PE*	*SH*
Panels with strong advantage for men	4	2	1	1
Panels with advantage for men	2	1		1
Neutral panels	9	5	4	
Panels with advantage for women	3		3	
Panels with strong advantage for women	4	1	2	1
Grants gained (PE) and lost (LS, SH) by women	0.2	3	-4.2	1.4
Share of all grants received by women		7%	11%	16%

Source: Van den Besselaar & Mom (forthcoming)

In the Life Sciences (LS) and the Social Sciences and Humanities (SH), there are more panels with bias against than in favor of women, and in the Physics and Engineering (PE) panels it is the opposite. Table 1 also shows the effects on the gender distribution of the grants. In LS, women should have received 3 more grants, and in SH 1.4 more. On the other hand, in PE women received 4 grants more than expected when taking merit into account. The numbers seem low, but as a share of all grants for women it is substantial: 7%, 16% and 11% in LS, PE and SH respectively.

Panel characteristics and panel decision making

The literature distinguishes several panel characteristics that may influence decision making, such as *panel composition* and *panel dynamics* (Lamont 2009; Thorngate et al. 2009, Olbrecht & Bornmann 2010; Van Arensbergen et al. 2014b). The latter we cannot test here, but the first we can, using available data about the panelists, such as gender, experience (first year member or more years membership) and nationality. Would more female panel members increase the probability for women applicants to be successful? This has been a leading idea behind gender equality policies in research councils, including the case studied here. Or does it work opposite, based on the so-called *queen bee* mechanism (Faniko et al. 2020): Women in the panel are experienced researchers that made a successful career in a male dominated science system. In order to be able to do that, female panelist may have internalized the same stereotypes as men have, and according to the queen bee theory, often even stronger than men.

As the council under study strongly emphasizes gender equality, more *experienced* panel members may have internalized this policy more than those who are in the panels for the first time. If so, one would expect that the higher the (average) experience of panelists, the lower the level of gender bias against women in that panel.

Another factor mentioned in the literature that influences decision making is *work-load*. The higher the work-load, the more decision making is based on heuristics instead of on a detailed assessment of the individual grant applications (Van Arensbergen et al. 2014). And these heuristics can be expected to reflect at least partly *gender stereotypes* panel members may have. The expectation would be that the higher the work-load, the stronger the role of gender stereotypes, and the higher the bias against women applicants.

Measuring gender stereotyping

How can we measure the strength of these stereotypes? Literature has shown that gender stereotypes influence spoken and written language, and that this can take different forms. (i) The same criteria may be used for men and women, but the evaluation of men tends to focus on the *presence* of required qualifications, whereas the evaluation of women tends to focus on the *absence* of qualifications (Uhlmann & Cohen 2005). One therefore would expect that men are evaluated in positive terms (presence of the required quality) and women in negative terms (absence). For example, independence is an important criterion in the grant instrument under study. In the evaluation of men, one would expect that independence is mentioned when the male applicant indeed is independent, but it is not mentioned if the male applicant is not independent. For women it would work in the opposite way. If present, independence is not mentioned, but if absent it would be mentioned as a negative point (Vinkenburg et al. 2014). (ii) Communication theory suggests that stereotyping generally leads to the use of negation words (like not, no, never) for the 'out group': negation bias (Beukenboom et al. 2010). Within a male dominated system, women are the outgroup, and consequently, positive characteristics of women are formulated in negative terms ("She is not bad" versus "he is good"). This would add to the use of negation terms in reviews about women. (iii) As the excellent scientist stereotype is male, male (agentic) traits such as being assertive, confident, ambitious, dominant are considered important (Abele & Wojciszke 2014). One would therefore expect that the stronger the gender stereotypes in a panel, the more agentic terms are used in review reports, and in combination with the previous points differently for men than for women. In review reports about men one would expect agentic terms to be used in a positive way, whereas in review reports of women one would expect that the use of agentic terms refer to the lack of those traits. So for women the frequency of agentic terms would correlate positive with gender bias *against* women, and for men one would expect that the frequency of agentic terms would correlate positively with bias *in favor* of men. (iv) Gender stereotyping depends on implicit and automatic attitudes and opinions of panel members, but other panels characteristics influence the role of stereotyping. For example, high work pressure may result in *decision heuristics* replacing the individual assessment of applicants (Van Arensbergen et al. 2014a), and those heuristics may bring in gender stereotyping.

Data & methods

To analyze the *effects of panel characteristics* on the level of bias, we use the following variables: (i) level of gender stereotyping, measured by the frequency of negation words, and of agentic words in the review reports – aggregated at the panel level; (ii) the number and share of female panel members; (iii) the diversity of panels in terms of countries represented in the panel; (iv) the experience of the panel members (number of years in the panel); and (v) the work load (number of applicants per panelist).

The Linguistic Inquiry and Word Count (LIWC) program (Tausczik & Pennebaker 2010; Seih et al. 2017) was used to produce the linguistic variables based on the review reports. The project proposals were processed with LIWC, using the 2015 English dictionary (Seih et al. 2017). We could use an existing dictionary for *agentic terms* (Madera et al. 2009), and the standard *negation words* category of LIWC. LIWC provides for each of the proposals the percentage of negation words and the percentage of agency words. The resulting percentages of the relevant linguistic categories have been averaged per panel for men and women separately.

An exploratory analysis gives the *correlation* between the gender bias indicator and several panel characteristics. This shows what panel characteristics predict gender bias.

Findings: Gender bias and panel characteristics

Why do the scores for gender bias differ between the panels? One possible interpretation is that random variation in the decision-making process will lead to variety of bias scores at the panel level, without a systematic pattern. At panel level, men and women as a group sometimes win and sometimes lose, but over a longer period at the aggregated level, there may be equality. This would also explain the contradictory findings between studies on gender bias in grant allocation.

Another possibility is that the differences in gender bias relate to panel characteristics, such as panel composition, processes, procedures, and to the opinions and implicit gender stereotypes of the panelists. If there is no gender bias, one would expect no systematic correlation between panel characteristics and the bias indicator. The number of panels (22) in this study does not enable an in-depth multivariate analysis, but it allows for an exploratory analysis.

Differences between the disciplinary domains.

Before analyzing the differences between the panels, we firstly address differences at the higher level of the disciplinary domains. In domains with many female applicants (Social sciences (SH) and in Life sciences (LS)), we found that bias tends to be against women, but in the domain with low percentages of women (Physics and Engineering (PE)) it tends to be the opposite. This is similar to the findings of Van der Lee and Ellemers (2015), who suggested that panelists in fields with many female researchers may become less aware that gender bias is still relevant, and (male and female) panelist may become much more critical towards women. On the other hand, in PE, women have more success than expected. An explanation could be that within these fields a more diverse staff (and student population) is needed and wanted, and panels may therefore have some preference for women when deciding about grants[c]. We indeed found a positive relation between the share of women applicants in a field and the level of gender bias (r = 0.197). If our argument is correct, then a consequence would be that if the share of women in the PE domain increases, the domain may become more similar to LS and SH fields – leading to a stronger overall bias against women.

Work load

The procedure in the studied funding instrument consists of two steps. In the first step we found bias against women, and in the second step the bias was in favor of women (Van den Besselaar & Mom, forthcoming). Small group research suggests that the higher *work-load* in a group, the more members will use heuristics (such as gender stereotypes) for decision-making, instead of assessing individuals on their merit (Olbrecht & Bornmann 2020; Van Arensbergen et al. 2014b). This may explain that in step 1 gender bias is in favor of men, as in that step work pressure is much higher for the panelists who have to assess a huge amount of applications in a short time. In step 2 this is different, as only 25% of the applicants make it to step 2, and consequently work load in step 2 is much lower than in step 1. However, one would also expect to find this pattern at the level of panels: do panels with higher numbers of applicants and so higher work load have a stronger bias against women? If this would be the case here, one would expect a positive correlation between the workload (number of applicants divided by the number of panelists) and the bias indicator. As Table 2 shows, this is not the case (r = - 0.214).

Table 2. Factors influencing gender bias at the panel level

Variables	*Pearson correlation with the bias indicator**	*Significance level*
Workload	-0.214	0.340
Share female applicants	0.290	0.191
Gender stereotyping		
Negation words (women)	0.601	0.003
Negation words (men)	0.350	0.110
Agentic words (review reports women)	0.458	0.032
Agentic words (review reports women)	0.277	0.211
Panel composition		
Number female applicants	0.195	0.385
Number recurring female applicants	0.054	0.812
Number recurring male applicants	-0.441	0.040
Country diversity	-0.460	0.031

* Positive bias score = bias in favor of women, negative bias scores = bias in favor of men. The higher the score, the more beneficial the panel is for women; the lower the score, the more beneficial it is for men.

Gender stereotyping

One of the most discussed causes of gender bias is *gender stereotyping*, which - as discussed earlier - can be identified through the analysis of language use. *Negation* terms in review reports not only indicate negative evaluations of applicants but also reflect gender stereotypes about women (Beukeboom et al 2010; Beukeboom 2014; Beukeboom & Burgers 2017). As gender stereotyping may lead to gender bias against women (Ellemers 2018; Varian 1999; Steward & Valian 2018), one would expect that the higher the amount of negation words in review reports about women (aggregated at the panel level), the higher gender bias score for that panel – which is indeed the case (r = 0.601, p = 0.003). The correlation between the gender bias indicator and negation words in review reports of men is also positive but lower (r = 0.350, p = 0.110), implying that in panels where review reports about men use more negation words, the level of bias in favor of men us higher (Table 2).

Agentic characteristics (Madera et al, 2009) are considered important for scientists. At the same time, men are expected to possess agentic characteristics, but for women this is often seen as an exception. Therefore, one would expect that women are differently assessed in terms of agentic terms than men. We indeed find at panel level a positive relation between the bias indicator and the use of agentic terms in reviews about men (r = 0.273, p = 0.211), implying that the more agentic terms in the review, the higher bias *in favor* of men. For women we find an even stronger correlation between the gender bias indicator and the use of agentic terms (r = 0.458, p = 0.032). So, the more agentic terms, the higher bias *against* women (and in favor of men). Obviously, when agentic characteristics are discussed, it is done in a positive way for men, but in a negative way for women.

Panel composition

Despite the often-used policy of increasing the number of female panelists, we found that the *more* women are in a panel, the higher the panel does score on gender bias (r = 0.195), which also was found in other studies (Border 1993). As discussed, this can be explained by the queen bee phenomenon.

Gender equality is a clear and visible priority within the council, and therefore one would expect that experienced reviewers (= not a first-year panel member) would be more aware of gender bias and act accordingly: a negative correlation between the average experience of panel members and gender bias. This indeed works for men (r = -0.441, p = 0.04) but not for women (r = - 0.054). Experienced women seem to remain critical on women applicants, also in line with the Queen Bee phenomenon (Faniko et al, 2020). Finally, diversity of a panel in terms of nationalities represented has a positive effect for women (r = -0.460, p = 0.031).

We were also able to test this partly also for other years (2009-2014), albeit without using the merit variables as covariates which we only have for 2014. A multiple linear regression was done to investigate the relation between *diversity in the panel* and the *success rate of female applicants*. As panels may differ in success rate, and female applicants are not evenly distributed over the panels, we have to control for the *overall success rate* of panels. We do this analysis for all panels we have this information for, that is 150 panels over the 2009-2014 period. Table 3 shows the results, which are similar to those for 2014 only: Controlling for the panels' overall success rate, the country diversity has a positive effect on the female success rate, but the share of female panel members in a panel has a significant negative effect.

Table 3. Success rate of female applicants by share of female panelists and by country diversity

	Standardized Beta	T	Sig.
(Constant)		-0.977	0.330
Share of female panel members	-0.151	-2.094	0.038
Number of countries in panel	0.122	1.71	0.089
Overall success rate	0.444	6.187	0.000

$R^2 = 0.25$

Conclusions and discussion

In our case, 27% of the panels there is bias *against* women, and in 32% there is gender bias *in favor* of women, leading to an about equal overall result. Women lose in LS and SH, but win in PE. Several panel characteristics were found to correlate with the gender bias against women, such as the share of female panelists (positive correlation), the share of experienced panelists, and the international variety of panel members (both negative correlation). Using a linguistic analysis of the review reports to measure gender stereotyping, we found that the higher the level of gender stereotyping at the panel level, the stronger gender bias against women.

This study has several *limitations:* The findings at the panel level are explorative. Although the initial results are promising, we would welcome a study based on a much larger sample of panels and funding instruments. In that way, it would be possible to systematically relate the level of gender bias with the relevant characteristics of panels and panel members within their organizational and possibly national context. Another limitation relates to the merit variables, as several others could be included representing a variety of other merit dimensions, such as the applicant's independence (van den Besselaar & Sandstrom 2019), and the number of awards and prizes.

Several lines of *further research* can be distinguished. First, women seem to apply for grants less often than men do, which leads to gender differences in winning research grants independently of the decision-making process. The role of self-selection within application behavior is therefore an interesting topic for further research. Second, several other forms of bias exist, based on nepotism, cognitive proximity, or ageism. These forms of bias may work out differently for men and women. Third, as the quality of the textual and (during the interview) the oral presentation of the proposal may play a role in the panel assessment, the question comes up whether men and women differ in terms of self-presentation, and whether this influences the selection process.

Acknowledgements

This paper is a result of the GendERC project, which received funding from the ERC under grant 610706. Contract partners were Vrije Universiteit Amsterdam (Amsterdam, Netherlands), Joanneum Research (Vienna, Austria), and Tecnalia (Bilbao, Spain). The authors are grateful to especially Lucia Polo (Tecnalia), who collected data about the panel members, and to Camiel Beukeboom (Vrije Universiteit Amsterdam) who provided useful advice on how to analyze linguistic bias.

References

Abele, A.E. & Wojciszke, B. (2014). Chapter four – Communal and agentic content in social cognition: a dual perspective model. *Advances in experimental social psychology,* 50, 195-255.

Beck, R. & Halloin, V. (2017). Gender and research funding success: Case of the Belgian F.R.S.-FNRS. *Research Evaluation,* 26 (2), 115–123.

Beukeboom, C.J., Finkenauer, C. & Wigboldus, D.H.J. (2010). The negation bias: when negations signal stereotypic expectancies. *Journal of Personality and Social Psychology,* 99, 978-992.

Beukeboom, C.J. (2014). Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Social cognition and communication,* 31, 313-330.

Beukeboom, C.J. & Burgers, C. (2017). Linguistic bias. *Oxford Research Encyclopedia of Communication*.

Bornmann, L., Mutz, R. & Daniel, H.-D. (2007). Gender differences in grant peer review: a meta-analysis. *Journal of Informetrics,* 1, 3, 226–238.

Broder, I.E. (1993). Review of NSF economics proposals: gender and institutional patterns. *American Economic Review,* 83, 964-970.

Ceci, S.J., Ginther, D.K., Kahn, S. & Williams, W.M. (2014). Women in academic science: a changing landscape. *Psychological Science in the Public Interest*, 15, 3, 75-141.

Ellemers, N. (2018). Gender Stereotypes. *Annual Review of Psychology,* 69, 275–298.

Faniko, K., Elllemers, N. & Derks, B. (2020). The Queen Bee phenomenon in academia 15 years later: does it still exist, and if so why? *British Journal of Social Psychology*, 60(2), 383-399.

Friessen, H. (1998). Equal opportunities in Canada. *Nature,* 391, 326.

Ginther, D. & Kahn, S. (2006). Does science promote women? Evidence from academia 1973-2001: in: *Science and engineering careers in the United States: an analysis of markets and employment,* 2009, 163-194.

Grant, J., Burden, S. & Breen, G. (1997). No evidence of sexism in peer review. *Nature,* 390, 438

Holman, L., Stuart-Fox, D. & Hauser, C.E. (2018). The gender gap in science: how long will it take until women are equally presented? *Plos Biology,* 16(4): e2004956

Hosek, S.D. (2005). *Is there gender bias in federal grant programs*? Accessed online August 21, 2020; https://www.rand.org/pubs/research_briefs/RB9147.html).

Jayasinghe, U.W., Marsh, H.W. & Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: the effect of assessor ad researcher attributes on assessors' ratings. *Journal of the Royal Statistical Society A,* 166, part 3, 279-300

Kahn, S., Ginther, D. (2018). Women and Science, Technology, Engineering and Mathematics (STEM): Are Differences in Education and Careers due to Stereotypes, Interests or Family?. In Averett, S., Argys, L.M., Hoffman, Saul D.H., eds. The *Oxford Handbook of Women and the Economy*. Oxford University Press. Online October 2017. Print June 2018

Lamont, M. (2009). *How professors think. Inside the curious world of academic judgement.* Harvard University Press.

Leslie, S.J., Cimpian, A., Meyer, M. & Freeland, E, (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science,* 347(6219), 262-265.

Ley, T.J. & Hamilton, B.H. (2008). The gender gap in NIH grant applications. *Science,* 322, 1472-1474.

Madera, G.M. Hebl, M.R. & Martin, R.C. (2009). Gender and letters of recommendation for academia: agentic and communal differences, *Journal of Applied Psychology,* 94, 1591–1599.

Marsh, H.W., Jayasinghe, U.W. & Bond, N.W. (2011). Gender differences in peer reviews of grant applications: A substantive-methodological synergy in support of the null hypothesis model, *Journal of Informetrics,* 5, 167–180.

Marsh, H.W., Bornmann, L, Mutz, R., Daniel, H.-D. & O'Mara A.O. (2009). Gender effects in the peer reviews. *Review of Educational Research,* 79, 1290-1326.

Marsh, H.W., Jayasinghe, U.W. & Bond, N.W. (2008). Improving the peer review process for grant application. *American Psychologist,* 63, 160-168.

Mood, C. (2009). Logistic regression: why we cannot do what we think we can do, and what we can do about it. *European Sociological Review,* 26, 67-82.

Seih, Y-T, Beier, S. & Pennebaker, J.W. (2017). Development and examination of the Linguistic Category Model in a computerized text analysis method. *Journal of Language and Social Psychology,* 36, 343-355.

Stewart, A.J. & Valian, V. (2018). *An inclusive academy, achieving diversity and excellence*. MIT press

Olbrecht, M. & Bornmann, L. (2010). Panel peer review of grant applications: What do we know from research in social psychology on judgment and decision-making in groups? *Research Evaluation,* 19 293–304.

Sandström, U. & Hällsten, M. (2008). Persistent nepotism in peer review. *Scientometrics,* 74(2) 175-189.

Sekaquaptewa, D., Espinoza, P., Thompson, M., Vargas, P. & Von Hippel, W. (2003). Stereotypic explanatory bias: implicit stereotyping as a predictor of discrimination. *Journal of Experimental Social Psychology,* 39 (1)*,* 75-82.

Tausczik, Y.R. & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology,* 29, 24-54.

Thorngate, W., Dawes, R.M. & Foddy, M. (2009). *Judging merit*. New York: Psychology Press.

Uhlmann, E. L. & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination, *Psychological Science,* 16, 474–480.

Van Arensbergen, P., van der Weijden, I. & van den Besselaar, P. (2014a). Different views on scholarly talent – what are the talents we are looking for in science? *Research Evaluation,* 23, 4, 273-284.

Van Arensbergen, P., van der Weijden, I. & van den Besselaar, P. (2014b). The selection of talent as a group process; a literature review on the dynamics of decision-making in grant panels. *Research Evaluation,* 23(4), 298-311.

van den Besselaar, P. & Sandström, U. (2019). Measuring researcher independence using bibliometric data: A proposal for a new performance indicator. *PLoS ONE* ,14(3), e0202712.

van den Besselaar, P., Schiffbaenker, H., Sandström, U. & Mom, C. (2018). Explaining gender bias in ERC grant selection. *STI 2018 Conference Proceedings*, 346-352.

van den Besselaar, P. & Mom, C. (forthcoming). Gender bias in grant allocation – mixed findings.

Van den Lee, R. & Ellemers, N. (2015). Gender contributes to personal research funding success in the Netherlands. *PNAS,* 112, 12349-12353.

Varian, V. (1999). *Why so slow? The advancement of women*. MIT press.

Vinkenburg, C.J., Jansen, P.G.W., Dries, N. & Pepermans, R. (2014). Arena: A critical conceptual framework of top management selection. *Group & Organization Management,* 39(1), 33-68. doi:10.1177/1059601113492846

Wennerås, C. & Wold, A. (1997). Nepotism and sexism in peer review. *Nature,* 387, 341-343.

Williams, W.M. & Ceci, S.J. (2011). Understanding current causes of women's underrepresentation in science. *PNAS,* 108(8), 3157-3162.

Witteman, H.O., Hendricks, M., Straus, S. & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *The Lancet,* 393, 531-540.

[a] Currently scientific advisor at DZHW, Schützenstraße 6a, 10117 Berlin (Germany)

[b] Three panels covering qualitative social sciences, humanities and law are excluded from this analysis, as the bibliometric data may not be valid for measuring performance in these panels.

[c] Research on selection of professors in these field suggests the same preference.

Crossref as a source of open bibliographic metadata

Nees Jan van Eck and Ludo Waltman

{ecknjpvan, waltmanlr}@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (the Netherlands)

Abstract

Several initiatives have been taken to promote the openly availability of bibliographic metadata of scholarly publications in Crossref. We present an up-to-date overview of the availability of six metadata elements in Crossref: reference lists, abstracts, ORCIDs, author affiliations, funding information, and license information. Our analysis shows that the availability of these metadata elements has improved over time. However, it also shows that many publishers need to make additional efforts to realize full openness of bibliographic metadata. To illustrate the value of open metadata, we use the metadata in Crossref to construct and visualize a large citation network of scholarly journals.

Introduction

Many scholarly publishers work together with Crossref (https://www.crossref.org/) to register Digital Object Identifiers (DOIs) for their publications. These publishers have the possibility to submit bibliographic metadata for their publications to Crossref. This metadata is then made openly available by Crossref, with the possible exception of the reference lists of publications, which may be kept closed. By making large amounts of bibliographic metadata openly available, Crossref is becoming an increasingly interesting data source for bibliometric analyses (Hendricks et al., 2020).

The Initiative for Open Citations (https://i4oc.org/), launched in 2017, has led to a major increase in the open availability of bibliographic references in Crossref, especially after Elsevier's recent decision to support the initiative (Plume, 2020; Waltman, 2020a). Likewise, the Initiative for Open Abstracts (https://i4oa.org/), launched in 2020, has contributed to an increase in the availability of abstracts in Crossref. Nevertheless, there are still many publications in Crossref for which the reference list or the abstract has not yet been made openly available. In addition, other metadata elements, such as ORCIDs and affiliations of authors and license and funding information, may not be available either.

In this paper, we present an up-to-date overview of the availability of bibliographic metadata in Crossref, focusing on six metadata elements: reference lists, abstracts, ORCIDs, author affiliations, funding information, and license information. We show how the availability of these metadata elements has improved over time (see also Habermann, 2019; Hendricks et al., 2020), and we analyze the contributions made by different publishers. We illustrate the value of open bibliographic metadata by using the metadata in Crossref to construct and visualize a large citation network of scholarly journals.

Data

We use Crossref's XML Metadata Plus Snapshot. The snapshot was downloaded on March 5, 2021. It includes the reference lists of publications in journals published by Elsevier and the American Chemical Society. These reference lists were opened in, respectively, January and February 2021. We consider only the 86.4 million records classified as journal article. For each article, we determined the year of publication. When a print year was provided, we used this year. When no print year was provided, we used another year field, typically the online year.

Availability of bibliographic metadata in Crossref

Time trend

Crossref includes a total of 52.8 million journal articles in the period 2000–2020. The annual number of articles has strongly increased during this period, from 1.3 million in 2000 to 4.7 million in 2020. Figure 1 shows how the availability of different metadata elements has improved over time. For each year in the period 2000–2020, the figure shows the percentage of articles that have an openly available reference list, an abstract, at least one ORCID, at least one author affiliation, funding information, and license information.

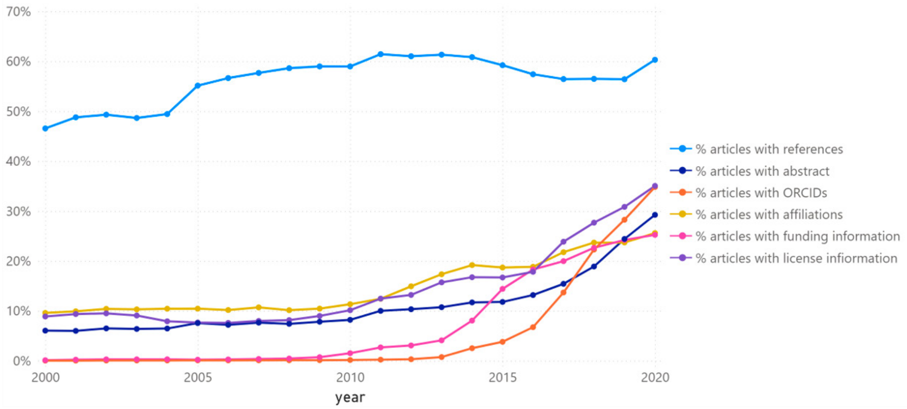


Figure 1. Availability of different metadata elements for journal articles in Crossref.

Thanks to the support publishers have given to the Initiative for Open Citations, reference lists have been made openly available not only for many recent articles but also for many older ones. As can be seen in Figure 1, the percentage of articles with an openly available reference list is fairly constant over time, starting just below 50% in 2000 and ending just above 60% in 2020. The percentage of articles for which an abstract is available has increased from 6% in 2000 to 29% in 2020. Most of the increase took place in recent years, thanks also to publishers' support for the Initiative for Open Abstracts.

In older years, there are almost no articles with ORCIDs. However, in recent years, the percentage of articles with ORCIDs has strongly increased, reaching 35% in 2020. This is an important development for bibliometric analyses in which the careers of researchers are traced. At the moment, such analyses typically rely on algorithms for author name disambiguation (Smalheiser & Torvik, 2009). In the future, when ORCIDs will have been widely adopted, the use of such algorithms may no longer be necessary.

The percentage of articles for which author affiliations are available has increased from 10% in 2000 to 26% in 2020. Affiliations are reported in unstructured strings, so they do not have a standardized format. In the near future, affiliations can be reported in a standardized way using Research Organization Registry (https://ror.org/) identifiers (Gould, 2020). Because this is a very recent development, we do not consider it in more detail in this paper.

Funding information is almost completely missing for older articles, but in recent years the availability of funding information has increased substantially. For 25% of the articles in 2020 funding information is available.

The percentage of articles for which license information is available has increased from 9% in 2000 to 35% in 2020. Crossref distinguishes between license information for three different versions of an article: The author's accepted manuscript, the version of record, and the version intended for text and data mining. In Figure 1, we consider license information only for the author's accepted manuscript and the version of record.

In the interpretation of the above statistics, it is important to be aware that all documents published in a journal are classified as journal article in Crossref. This includes not only research articles, but for instance also letters, editorials, book reviews, and corrections. Some of the documents published in a journal may not have a reference list, an abstract, or funding information. The percentage of journal articles for which all metadata elements are available will therefore never reach 100%.

Publishers

Different publishers handle the submission of bibliographic metadata to Crossref in different ways. Some publishers submit as much metadata as possible, while other publishers submit only specific metadata elements. The latter publishers may prefer not to submit certain metadata elements, such as abstracts or author affiliations, to Crossref. It is also possible that they are not aware of the possibility to submit these metadata elements to Crossref or that they do not have the technical expertise needed to submit them. In the case of reference lists, publishers may also choose to submit them to Crossref but not to make them openly available.

For the 8.8 million journal articles in Crossref in 2019 and 2020, we determined for each publisher the percentage of articles that have an openly available reference list, an abstract, at least one ORCID, at least one author affiliation, funding information, and license information. The resulting statistics are available in Zenodo (Van Eck & Waltman, 2021).

Almost all of the larger publishers support the Initiative for Open Citations, as can be seen on the website of the initiative (https://i4oc.org/). Our statistics indeed confirm that almost all of the larger publishers make the reference lists of their articles openly available in Crossref. The most important exception is IEEE, which submits reference lists to Crossref but does not make them openly available.

The Initiative for Open Abstracts is supported by a substantial number of larger publishers. However, the four largest publishers (i.e., Elsevier, Springer Nature, Wiley, and Taylor & Francis) do not yet support the initiative, even though Springer Nature does submit abstracts to Crossref for some of its articles. Our statistics indeed show that a number of larger publishers do not yet make abstracts available in Crossref. Similar statistics are also presented on the website of the Initiative for Open Abstracts (https://i4oa.org/).

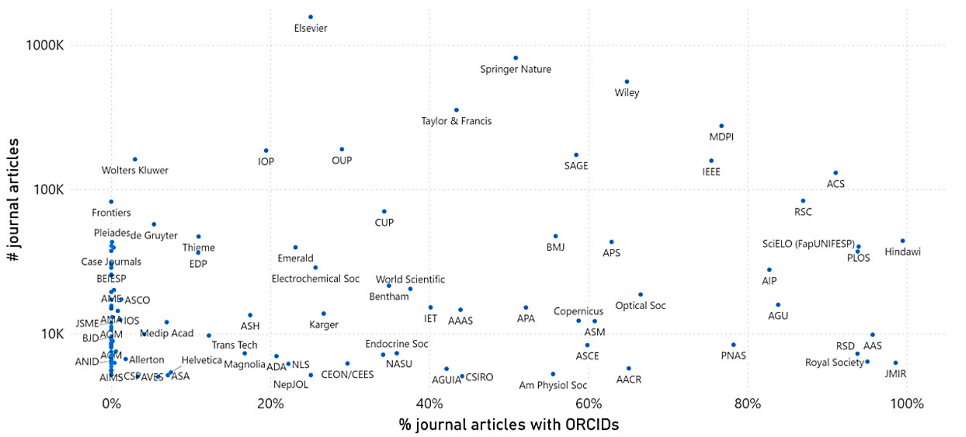


Figure 2. Number of journal articles of a publisher in 2019 and 2020 and percentage of articles with at least one ORCID.

For publishers with at least 5000 articles in 2019 and 2020, Figure 2 shows the percentage of articles with at least one ORCID. All publishers with more than 100,000 articles make ORCIDs available in Crossref. However, for some of these publishers, in particular Wolters Kluwer, the percentage of articles with ORCIDs is still quite low, while for others, such as the American Chemical Society, MDPI, and IEEE, it is already fairly close to 100%.

Figure 3 presents statistics for author affiliations. There is a clear separation between publishers that do not make affiliations available in Crossref and publishers for which almost all articles in 2019 and 2020 have affiliations in Crossref. Looking at the largest publishers, Figure 3 shows that most articles of Wiley and Taylor & Francis have affiliations, while affiliations are completely missing for articles of Elsevier, Springer Nature, and MDPI.

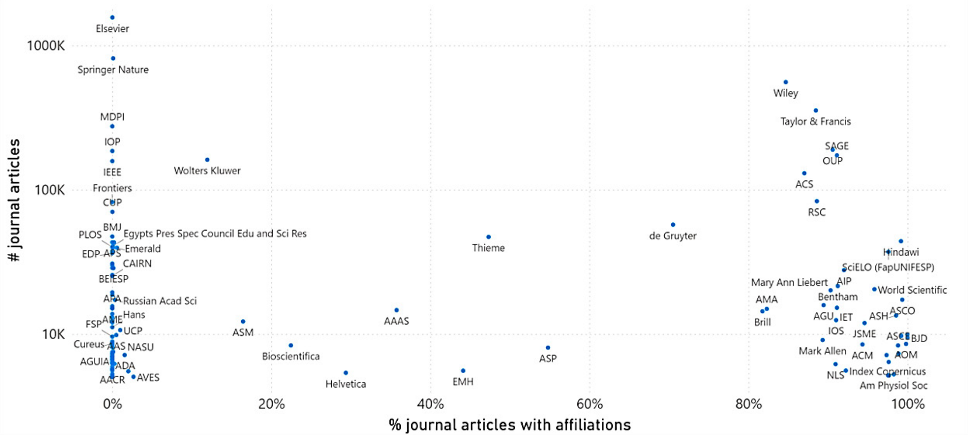


Figure 3. Number of journal articles of a publisher in 2019 and 2020 and percentage of articles with at least one author affiliation.

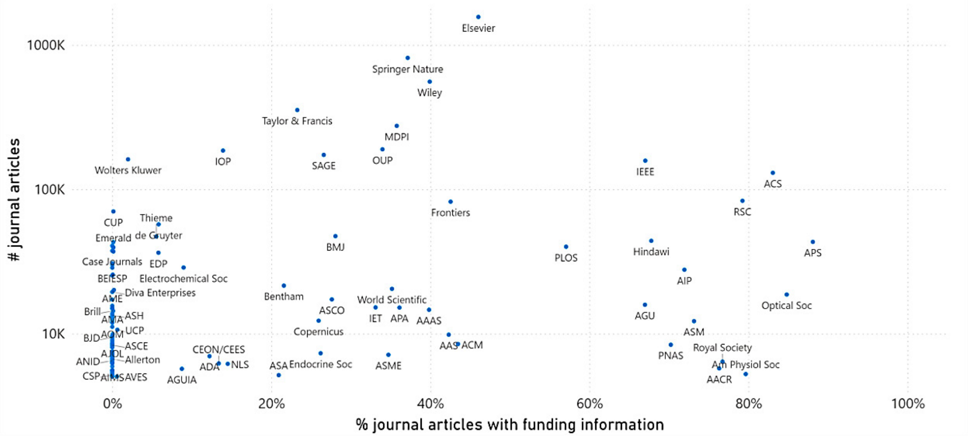


Figure 4. Number of journal articles of a publisher in 2019 and 2020 and percentage of articles with funding information.

As can be seen in Figure 4, all publishers with more than 100,000 articles in 2019 and 2020 make funding information available in Crossref, although for some of them the percentage of articles with funding information is quite low. Differences between publishers in the percentage of articles with funding information may be partly due to differences in the disciplinary profiles of publishers. For articles in the natural sciences it is more common to acknowledge funding than for articles in the social sciences and humanities. Publishers that are active mainly in the natural sciences can therefore be expected to have a higher percentage of articles with funding information than publishers that focus primarily on the social sciences and humanities.

Figure 5 shows that many publishers, including large ones such as Wiley and MDPI, have made license information available for all or almost all their articles in 2019 and 2020. However, many other publishers, including Elsevier, Springer Nature, and Taylor & Francis, have made license information available for only a small share of their articles, or they have not made any license information available at all.

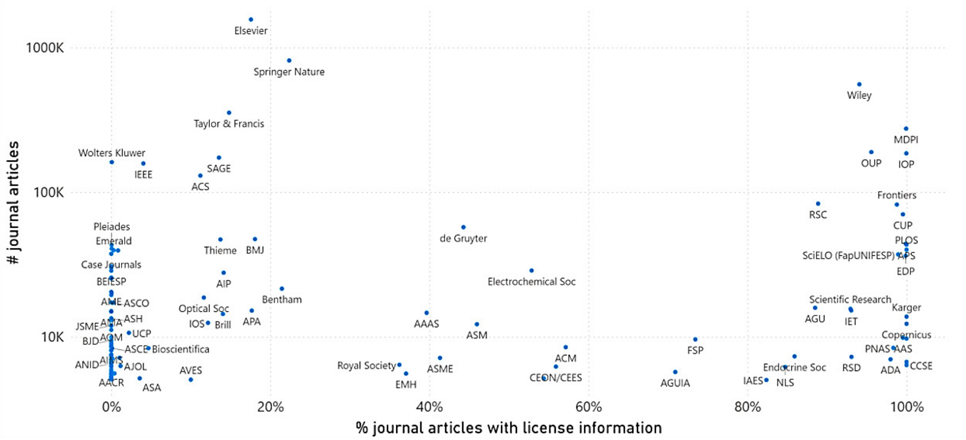


Figure 5. Number of journal articles of a publisher in 2019 and 2020 and percentage of articles with license information.

Constructing and visualizing a journal citation network

To illustrate the value of the open bibliographic metadata, we use the metadata in Crossref to construct and visualize a large citation network of scholarly journals.

Based on openly available reference lists, we identified 0.94 billion citation links between the 86.4 million journal articles in our data. We excluded all journals that have fewer than 300 articles with references to other articles in our data. Of the remaining 14,931 journals, we selected the 10,000 journals with the largest number of (incoming and outgoing) citation links with other journals. We constructed a citation network for the selected journals. This network includes 0.75 billion citation links between the selected journals. Using the VOSviewer software (Van Eck & Waltman, 2010; https://www.vosviewer.com/), we created a visualization of our journal citation network. The visualization can be explored interactively at https://bit.ly/3dNHA2A.

A few years ago we created a similar visualization (Van Eck & Waltman, 2017). The number of journals with openly available reference lists in Crossref was more limited at that time, and the visualization therefore included a smaller number of journals. However, the overall structure of the two visualizations is quite similar. This structure resembles the so-called consensus map of science discussed by Klavans and Boyack (2009).

Conclusion

Openness of bibliographic metadata is as an essential element in the broader development toward openness in scholarly publishing (Waltman, 2020b). Open bibliographic metadata helps to make bibliometric analyses more transparent, reproducible, and inclusive. It also helps to make it easier for researchers and others to find the most relevant scholarly literature.

We have shown that the open availability of different metadata elements in Crossref has improved over time (see also Habermann, 2019; Hendricks et al., 2020). The increasing availability of reference lists, abstracts, ORCIDs, author affiliations, funding information, and license information is an important development. However, many publishers need to make additional efforts to realize full openness of bibliographic metadata. Publishers often do a good job in making certain metadata elements openly available, but they fail to do the same for other metadata elements. For instance, Wiley is among the leading publishers in terms of the availability of ORCIDs, author affiliations, and license information in Crossref, but it has not made any abstracts openly available. We hope that the statistics presented in this paper will help publishers to move toward full openness of bibliographic metadata.

A more in-depth analysis is currently in preparation. We plan to calculate more detailed statistics on the availability of different metadata elements in Crossref. In addition to journal articles, other content types, such as articles in conference proceedings, books and book chapters, preprints, data sets, and peer review reports, will be considered as well. We also plan to analyze the availability of links between different content types, for instance between journal articles on the one hand and preprints, data sets, and peer review reports on the other hand.

Acknowledgments

We are grateful to our CWTS colleagues Henri de Winter and Patrick Kooij for their help in processing the Crossref data. We thank Bianca Kramer and Catriona MacCallum for helpful feedback on an earlier version of this paper.

References

Gould, M. (2020). Publishers, are you ready to ROR? *Crossref.* https://www.crossref.org/blog/publishers-are-you-ready-to-ror/

Habermann, T. (2019). The big picture - Has CrossRef metadata completeness improved? *Metadata Game Changers*. https://metadatagamechangers.com/blog/2019/3/25/the-big-picture-how-has-crossref-metadata-completeness-improved

Hendricks, G., Tkaczyk, D., Lin, J. & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1), 414–427. https://doi.org/10.1162/qss_a_00022

Klavans, R. & Boyack, K.W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, 60(3), 455–476. https://doi.org/10.1002/asi.20991

Plume, A. (2020). Advancing responsible research assessment. *Elsevier*. https://www.elsevier.com/connect/advancing-responsible-research-assessment

Smalheiser, N.R. & Torvik, V.I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, 43, 1–43. https://doi.org/10.1002/aris.2009.1440430113

Van Eck, N.J. & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Van Eck, N.J. & Waltman, L. (2017). Visualizing freely available citation data using VOSviewer. *CWTS blog*. https://www.cwts.nl/blog?article=n-r2r294

Van Eck, N.J. & Waltman, L. (2021). *Crossref metadata statistics* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4719469

Waltman, L. (2020a). Q&A about Elsevier's decision to open its citations. *Leiden Madtrics*. https://leidenmadtrics.nl/articles/q-a-about-elseviers-decision-to-open-its-citation

Waltman, L. (2020b). Publications should be FAIR. *Leiden Madtrics*. https://www.leidenmadtrics.nl/articles/publications-should-be-fair

The Uptake of a Label for Peer-reviewed Books: the Flemish GPRC label

Eline Vandewalle[1], Raf Guns[1] and Tim Engels[1]

[1] *eline.vandewalle@uantwerpen.be, raf.guns@uantwerpen.be , tim.engels@uantwerpen.be*
University of Antwerp, Faculty of Social Sciences, Centre for R&D Monitoring (ECOOM) Middelheimlaan 1,
2020 Antwerp (Belgium)

Abstract

This research in progress paper provides preliminary results of an analysis of the uptake of the Flemish GPRC label for peer-reviewed books by scholars in the Social Sciences and Humanities (SSH). The GPRC label (Guaranteed Peer Reviewed Content, see www.gprc.be; Verleysen & Engels, 2013) was adopted in 2010 and is being used in the context of the regional performance based university research funding system. In the period 2010 to 2018 a total of 2312 GPRC-labelled publications have been included in the VABB-SHW, a regional database for SSH publications. Our analysis focuses on the uptake of the label across disciplines, the use of Dutch, English and other languages and the type of publications (monographs, book chapters, edited volumes and proceedings papers). The results show that the GPRC label is particularly relevant for the disciplines of Criminology and Law and for publications in Dutch.

Introduction

In recent years, the problem of how to include books in national databases and evaluation systems has generated some attention. The importance of books to many Social Sciences and Humanities (SSH) disciplines in conjunction with the difficulty of identifying peer-reviewed books and assessing their quality, has prompted several European countries to adopt new strategies for 'taking books into account' in national research evaluation systems, for the allocation of funding and for the creation of comprehensive national databases (see Giménez-Toledo et al., 2016; Giménez-Toledo et al., 2019*)*.

In Flanders, the GPRC label was created to indicate the peer-review status of books, particularly in the SSH. The GPRC label (Guaranteed Peer Reviewed Content label, see www.gprc.be; Verleysen & Engels, 2013) was created by the Flemish Publishers' Association to enable academic books and chapters published by Flemish publishers to be included in the VABB-SHW (henceforth VABB), a Flemish regional database for the SSH used in the context of the Flemish performance-based research funding system (PRFS). The rationale behind the creation of the label was to increase the coverage of books in the VABB and to enable local publishers with a hybrid portfolio of peer-reviewed books and books that are not peer-reviewed to have their peer-reviewed content included in the VABB. The GPRC label allows publishers to indicate that a book has been subjected to peer review prior to publication. The GPRC label can be seen as a testcase for the inclusion of books on an individual level. The potential wider international relevance of the label was mentioned by Verleysen and Engels, who argue that the development of similar labels in other countries could facilitate the coverage of SSH books in international databases and citation indexes (Verleysen & Engels, 2013). Meanwhile, a similar label has been created in Finland (see Giménez-Toledo et al., 2016 and www.tsv.fi/tunnus). Apart from the GPRC label, books can also be included in the VABB based on the publisher and by book series. This last option is more recent, and becoming more important as more book series are added.

We use data available in the VABB database to document the use of the label from its inception in 2010 until 2018. Earlier publications have explained the rationale for the creation of the GPRC label (Verleysen & Engels, 2013) and its potential pitfalls (Borghart, 2013), and have compared it to open-identity labels in book publishing (Kulczycki et al 2019). The current study documents the use of the GPRC label between 2010 and 2018 and takes a closer look at the language of GPRC publications as well as the distribution of the label among SSH disciplines.

The creation of the GPRC label

In Flanders, the size of the research fund universities can allocate at their own discretion is determined by the BOF-key, which takes into account several aspects of research performance including the number of peer-reviewed publications (see Engels & Guns, 2018). While the original BOF-key relied heavily on inclusion in the Web of Science indexes for all disciplines, public debate about the lack of coverage of the SSH disciplines in the Web of Science indexes prompted the Flemish government to provide funding to the Centre for R&D Monitoring (ECOOM) for the creation of the VABB database to cover the publications of scholars in the SSH more comprehensively. In order to be eligible for inclusion in the VABB, publications must (1) be publicly accessible, (2) be unambiguously identifiable by an ISSN or ISBN, (3) contribute to the development of new insights or the application thereof, (4) be peer-reviewed by independent experts in the field prior to publication and (5) consist of at least four pages (see Engels & Guns, 2018).

Bibliographic information on the publications in the SSH is submitted yearly by each Flemish university. The Authoritative Panel (GP), a panel of 18 eminent SSH scholars, decides which publication channels used by the SSH researchers who are affiliated to a Flemish university in previous years are included in the VABB. Their decision is informed by the criteria mentioned above. The lack of a consensus on what peer review for books is and how to evaluate whether a book has been subjected to peer review, is one of the major difficulties for including books in research evaluation systems (Verleysen & Engels, 2013). Moreover, book publishers often have hybrid portfolios of academic books directed at peers, textbooks for students and books for a wider audience. This makes the identification of a peer review process on the level of the publication channel -in this case the book publisher- problematic. Initially, only a limited set of prestigious publishers were automatically included in the VABB (Engels et al., 2012). The GPRC label focuses on the fourth criterion for inclusion in the VABB: peer review. In order to qualify for a GPRC label, the publisher has to hold on to a dossier that demonstrates the peer review process of the book. By introducing the requirement of a peer review dossier, the GPRC label has formalized and harmonized the peer review process of SSH book publications at Flemish publishers.

Publication patterns in the SSH

Several studies have analyzed publication patterns in the SSH in European countries (see Engels et al., 2012; Engels et al., 2018; Kulczycki et al., 2018; Kulczycki et al. 2020). The following section introduces a few general trends of publication patterns in the SSH in Flanders and internationally as well as some hypotheses about the uptake of the GPRC label.

Firstly, studies into publication patterns in non-English speaking European countries have looked at the use of English in academic publications. Kulczycki et al. (2018) reported an increase in the use of English in SSH publications in eight different countries, including Flanders. For Flanders, a steady rise of the percentage of publications in English with a corresponding drop in the share of Dutch language publications and publications in other languages is evident (Guns & Engels, n.d). However, there are large differences between SSH disciplines. While the majority of publications are written in English in most SSH disciplines, disciplines such as Law and Criminology still publish mostly in Dutch. The use of English in publications is associated with an international orientation while the use of Dutch indicates a local audience.

In terms of publication type, book publications make up about 21% of all VABB publications (Guns & Engels, n.d.). The VABB distinguishes between three types of book publications: chapters in books, monographs and edited volumes. The largest share of book publications, 16% of all publications are book chapters. In an analysis of publication practices in the Flemish SSH disciplines, 'no overall shift towards the journal article as the chosen publication vehicle is observed' (Engels et al., 2012, p. 387). Again, there are differences between the disciplines. Considering the differences between the Humanities and the Social Sciences, the Humanities, on the whole, publish more books and write more frequently in Dutch (Engels et al., 2012). It is to be expected that the Humanities, as relatively more book-oriented and locally oriented disciplines, use the GPRC label more often than the Social Sciences. In an international context, Kulczycki et al. (2018) find different evolutions of book publications in different European countries. While there is a relatively stable share of book chapters and monographs for Flanders, in other countries, most notably Poland, the share of book publications has dropped considerably because of changes in research policy (Kulczycki et al., 2018, p. 475).

While earlier studies reported on a rapid increase in the number of publications included in the VABB (Engels et al., 2012), the latest figures on the VABB show a decline in the number of publications for the past few years. Looking at the data on publications across all types, there has been a drop in the number of articles in journals and proceedings papers as well as book chapters (see Guns & Engels, n.d.).

Based on the previous studies about publication patterns in the SSH, we put forward three hypotheses. Firstly, the continued importance of books for SSH disciplines coupled with the lack of (non-English language) books in commercial databases might have created an impetus for the use of the label. We therefore expect to see the use of the label increase over the time period. Secondly, the use of the label may differ between disciplines. Some disciplines are more book-oriented than others. Moreover, the GPRC label has a local focus, only including books by local publishers, and thus might show a higher share of Dutch language publications and a higher share of publications among disciplines with a local focus.

Description of the data

Uptake of the label

The total number of publications in the dataset is 2312. Of these, 257 are monographs, 321 are edited volumes, 1717 are book chapters and 17 are proceedings papers. Until 2015, the use of the GPRC label increased steadily (Figure 1). However, the data show a significant drop in the use of the label since 2015, when the number of GPRC publications reached a maximum of 421. Among GPRC publications, the drop is most clearly visible among book chapters and edited volumes. As mentioned above, there has been a general decline in the number of publications in the VABB in the past couple of years. The number of GPRC-labelled monographs has fluctuated in recent years.

Language

Contrary to other types of VABB publications, GPRC publications are predominantly written in Dutch (68,3 %) and only to a lesser degree in English (28,6 %). This may indicate that book publications are more often targeted towards local audiences, especially when they are written for 'enlightenment purposes' (Engels et al., 2018, p. 595; Hicks, 2005). Another explanation for the larger share of publications in Dutch is the fact that the GPRC label can only be used by local publishers. Still, almost 30% of GPRC-labelled books are written in English. Only around three percent of publications are written in languages other than Dutch and English. They are mainly written in French or German, the other national languages of Belgium.

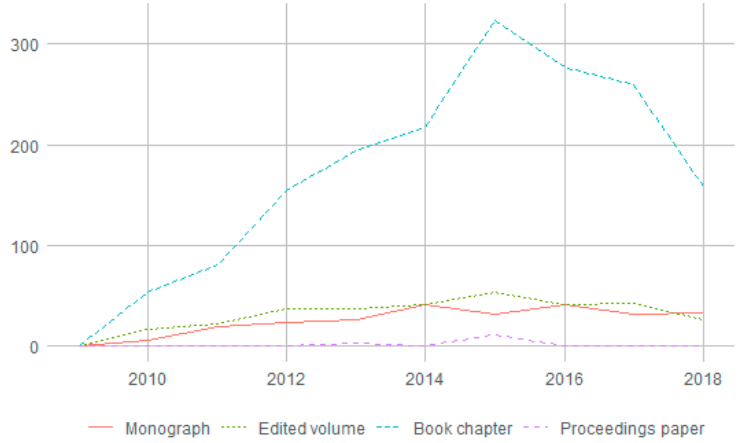


Figure 1: The evolution of the number of GPRC publications by year (by type).

Discipline

In terms of disciplines, we used the organizational classification of the VABB, which links each publication to disciplines based on the institutional affiliation of the researchers. In the VABB classification system, there are nine Humanities disciplines, seven Social Sciences disciplines and two general categories (Humanities-general, Social Sciences-general). For this analysis, whole counting was used: publications that were assigned to multiple disciplines were counted for each of the disciplines. As can be seen in Table 1, the discipline Law counts the highest number of GPRC publications in absolute terms- more than twice as many GPRC publications as the next discipline. Law has a local orientation. It has the largest share of publications in Dutch, followed by Criminology. Criminology takes the lead when the size of the disciplines is taken into account. Apart from Criminology and Law, and not taking into account the general categories, the other disciplines that use the GPRC label most often relative to their size are Sociology, Art History and History. This seems to corroborate the idea that disciplines with a local orientation would use the label more frequently. The disciplines that appear low on the list are Psychology, Social Health Sciences and Theology. Clearly, the GPRC label has not been taken up by all disciplines equally. Psychology, one of the larger SS disciplines, accounts for only 39 GPRC-labelled publications. Economics & Business, roughly as large as Law, also accounts for a relatively low number of GPRC publications (only 145). These disciplines are also known to be quite internationally oriented, with a large share of publications in English and a bigger presence in the Web of Science. While it is to be expected that more locally oriented disciplines such as Law and Criminology, but also History and Sociology use the local book label more frequently, the discrepancy between Criminology and Law and all other disciplines is striking and warrants further analysis. One important determining factor to be studied in particular is the influence of the location of the publisher and to what extent researchers publish in the local market (Flanders). A first look at the data showed that for the disciplines Law and Criminology, the GPRC-publishers are also the most used publishers for the discipline, while for the discipline History, this is not the case. This aspect will be explored further in the complete study.

Table 1: The distribution across disciplines ordered by share of GPRC publications

Disciplines	*Mono-graphs*	*Edited Volumes*	*Chapters, proceedings*	*All GPRC*	*Total VABB*	*GPRC/ VABB*
Criminology	23	79	294	396	2.590	15.29
Law	82	118	688	888	10.914	8.14
Sociology	16	27	208	251	4.727	5.31
Art History	33	28	146	207	4.032	5.13
History	44	33	133	210	4.184	5.02
Literature	15	33	106	154	3.479	4.43
Philosophy	27	26	115	168	5.233	3.21
Political Sciences	16	11	92	119	3.988	2.98
Archaeology	12	7	35	54	1.816	2.97
Humanities (General)	44	46	213	303	10.377	2.92
Linguistics	22	31	129	182	7.223	2.52
Communication Studies	11	7	65	83	3.358	2.47
Educational Sciences	20	9	76	105	4.330	2.42
Social Sciences (General)	24	35	183	242	12.250	1.98
Economics & Business	24	30	91	145	11.081	1.31
Theology	3	0	10	13	2.252	0.58
Social Health Sciences	5	6	43	54	12.208	0.44
Psychology	9	3	27	39	9.062	0.43

Table 1 also shows the distribution between the different publication types. It is possible that some disciplines tend to publish more monographs while other disciplines publish more book chapters and edited volumes. However, because the total amount of GPRC-labelled monographs is quite small (only 257), it is difficult to make statements about the differences between disciplines. Many SSH disciplines published roughly the same number of monographs. The highest count is found for law, with 82 publications. However, the second disciplines for GPRC monographs are History and Humanities General. As mentioned by Verleysen & Engels (2012), monographs are important to historians.

Conclusion

The GPRC label has enabled books published by local publishers to be included in the VABB database. It has also been a way for SSH researchers to have their locally published peer reviewed books be counted in the performance-based research funding system. Since its introduction in 2010, there has been an increase in the use of the label. However, since 2015, the use of the label has declined. We are currently analyzing the data in more detail to understand why this has happened.

While the GPRC label is open to all SSH researchers, some disciplines have used the label a lot more than others. The disciplines Criminology and Law take the lead in terms of number and share of GPRC-labelled publications. A discipline that appears to publish a lot of GPRC-labelled monographs relative to its size, is History. In terms of language, GPRC publications are written mostly in Dutch, this underwrites the idea that books are more often targeted towards a local audience, but also indicates that the GPRC label has been beneficial to the inclusion of non-English language books in the VABB. The location of the publisher could also have an effect on the use of the label.

The following parts of this study will take into account the publishers of GPRC-labelled books as well as the characteristics of GPRC publications in terms of authorship and the location of the publishers. We will also look at the alternatives to the GPRC label and the effect the GPRC label has had on the peer review practices among book publishers and conduct an in-depth interview with a member of the Authoritative Panel to get their view on the uptake of the label and the quality of the peer review dossiers submitted by publishers. Lastly, we aim to look at the GPRC label in the larger context of the SSH publications in the VABB and study whether the GPRC label has had an effect on the inclusion rate of books in the VABB.

References

Borghart, P. (2013). A label for peer-reviewed books? Some critical reflections. *Learned Publishing,* 26(3).167-171. https://doi.org/10.1087/20130303

Engels, T. C. E, Ossenblok, T. L. B., Spruyt, E. H. J. (2012). Changing publication patterns in the Social Sciences and Humanities, 2000–2009. *Scientometrics,* 93(2), 373-390. https://doi.org/10.1007/s11192-012-0680-2

Engels, T.C.E., Guns, R (2018). The Flemish performance-based research system: a unique variant of the Norwegian model. *Journal of Data and Information Science,* 3(4), 45-60. https://doi.org/10.2478/jdis-2018-0020

Giménez-Toledo, E., Mañana-Rodriguez, J., Engels, T., Guns, R., Kulczycki, E., Ochsner, M., Pölönen, J., Sivertsen, G., Zuccala, A. A. (2019). Taking scholarly books into account, part II: a comparison of 19 European countries in evaluation and funding. *Scientometrics,* 118(1), 233-251. https://doi.org/10.1007/s11192-018-2956-7

Giménez-Toledo, E., Mañana-Rodriguez, J., Engels, T., Ingwersen, P., Pölönen, J., Sivertsen, G., Verleysen, F. T., Zuccala, A. A. (2016). Taking scholarly books into account. current developments in five European countries. *Scientometrics,* 107(2), 685-699.

GPRC. (n.d.). Retrieved February 13, 2021 from https://www.gprc.be/

Guns, R., & Engels, T. C. E. (n.d*.). 4.2 Bibliometrische analyse van sociale en humane wetenschappen | Vlaams Indicatorenboek*. vlaamsndicatorenboek.be. Retrieved Januari 11, 2021 from: https://www.vlaamsindicatorenboek.be/4.2/bibliometrische-analyse-van-sociale-en-humane-wetenschappen

Hicks, D. (2005). The Four Literatures of Social Science. In H.F. Moed, W. Glänzel & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 473-496). Kluwer Academic Publishers. https://doi.org/10.1007/1-4020-2755-9_22

Kulczycki, E., Engels, T. C. E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., Nowotniak, R., Petr, M., Sivertsen, G., Istenič Starčič, A., Zuccala, A. (2018). Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics,* 116 (1), 463-486. https://doi.org/10.1007/s11192-018-2711-0

Kulczycki, E., Guns, R., Pölönen, J., Engels, T. C. E., Rozkosz, E. A., Zuccala, A. A., Bruun, K., Eskola, O., Starčič, A. I., Petr, M., & Sivertsen, G. (2020). Multilingual publishing in the social sciences and humanities: A seven-country European study. *Journal of the Association for Information Science and Technology*, 71(11), 1371–1385. https://doi.org/10.1002/asi.24336

Kulczycki, E., Rozkosz, E. A., Engels, T. C., Guns, R., Holowiecki, M., Pölönen, J. (2019). How to identify peer-reviewed publications: Open-identity labels in scholarly book publishing. *PloS One*, 14(3), e0214423. https://doi.org/10.1371/journal.pone.0214423

Label for peer-reviewed scholarly publications. (2015, May 27). Tsv.Fi. Retrieved April 20, 2021 from https://www.tsv.fi/en/services/label-for-peer-reviewed-scholarly-publications

Verleysen, F. T. , Engels, T. C. E. (2013). A label for peer-reviewed books. *Journal of the American Society for Information Science and Technology*, 64(2), 428-430. https://doi.org/10.1002/asi.22836

Verleysen, F., Engels, T. (2012). Historical publications at Flemish universities, 2000-2009. *Belgisch tijdschrift voor nieuwste geschiedenis,* 42(2), 110-143.

DISTINGUISHING SCIENCE COMMUNICATION & POPULARIZATION FROM RESEARCH-BASED PUBLIC INTERVENTIONS

Assessing societal impact of university research using written press documents

Florian Vanlee[1], Walter Ysebaert[2] and Hans Jonker[3]

[1] *florian.hendrik.j.vanlee@vub.be*

ECOOM-VUB, R&D Department, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Elsene (Belgium)

[2] *walter.ysebaert@vub.be*

ECOOM-VUB, R&D Department, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Elsene (Belgium)

[3] *hans.jonker@vub.be*

ECOOM-VUB, R&D Department, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Elsene (Belgium)

Abstract

Integrating the societal impact of university research (SIUR) into research policy and evaluation is increasingly advocated – based on the conviction that traditional conceptions of academic quality insufficiently address how research can excel. Calls for the expansion of objects and activities considered in assessment models have led to the development of new methods. Altmetrics, for instance, track the "popular" dissemination of published outcomes, while methods to map impact pathways identify the productive interactions researchers have with societal stakeholders to evaluate SIUR. Balancing feasibility with precision remains an issue, and existing methods tend to fall short in one of them. Hence, there is need for methodological expansion.

We demonstrate how presence in the written press offers a partial indication of SIUR. Discussing a case study conducted in Flanders, the Dutch-speaking Northern region of Belgium, we first demonstrate the objects referred to by the written press in relation to university research. Subsequently, we distinguish between coverage of, operationalizations of and interventions by research in the public debate. Based on this taxonomy, we argue that an extension of altmetrics to press coverage has potential, but operationalization must differentiate between popular attention for research and its outcomes and active research-based contributions to public debates.

Assessing the societal impact of university research (SIUR): Contribution and attribution

Increasingly, research quality assessment recognizes the legitimacy of integrating the societal impact of university research (SIUR) in evaluation efforts (Chowdhury, Koya, & Philipson, 2016). Whereas traditional evaluation models prioritize scholarly quality – measured with bibliometric indicators – recent years have witnessed growing calls to move beyond peer-reviewed publications as the primary modality to determine academic merit (Benneworth & Olmos-Peñuela, 2018; Chowdhury et al., 2016; Donovan, 2011; Hill, 2016). Policy attention for such calls is reflected to an extent by the introduction of indicators that aim to monitor and compare economic impacts of university research – often with patents and spin-offs as objects of analysis (Agrawal & Henderson, 2002; Czarnitzki, Rammer, & Toole, 2014). But even though (positive) economic effects of university research can be satisfactorily charted, they offer at best partial indications of SIUR (Bornmann, 2013). In fact, SIUR is generally distinguished from the economic impacts brought by university research, primarily to avoid its reduction to simplistic econometric assessment models (Hill, 2016). Hence, when and where SIUR is integrated in performance assessment models and allocation mechanisms – most notably in the UK, Australia and the Netherlands (Heyeres, Tsey, Yang, Yan, & Jiang, 2019; Rijcke, Wouters, Rushforth, Franssen, & Hammarfelt, 2016; Watermeyer & Hedgecoe, 2019), it tends to be defined in rather broad and generic terms. Insofar a certain degree of consensus has formed to contemporary perspectives on SIUR in scholarship and governance: SIUR cannot be narrowed down solely to academic impact nor economic impact. The consensus is thus negatively formulated.

Existing SIUR assessment models reflect the expansiveness of this negative conceptualization. Although traditional tools to gauge academic excellence face growing criticism for their reductive nature (Donovan, 2019) and harmful effects on science and its practitioners (Fochler, Felt, & Muller, 2016), they benefit from a degree of clarity and operability not yet available to SIUR evaluation practices. For instance, it is not hard to see how using bibliometric analysis to determine academic excellence is less demanding for parties involved (e.g. researchers, reviewers, funding bodies) than qualitative assessments based on impact narratives – like those required by the UK's Research Excellence Framework (REF) (Watermeyer & Hedgecoe, 2019). Indeed, integrating SIUR in research evaluation models clearly necessitates adopting decidedly more labour-intensive techniques and tools than those in place to measure academic or economic impact (Sivertsen & Meijer, 2020). The need to take a qualitative approach to SIUR– taxing as this may be – is based in no small part on the recognition that societal impact can rarely be causally attributed (Matt, Gaunand, Joly, & Colinet, 2017). Instead, SIUR emerges primarily through organic and contingent contributions to the agendas of societal stakeholders, as well as broader efforts to address societal challenges or even less defined segments of society at large (Robinson-Garcia, van Leeuwen, & Rafols, 2018). In fact, in many cases it is fairly implausible for SIUR to be an intentional, calculable outcome (Sivertsen & Meijer, 2020).

Consequently, scholarship on SIUR has explored in recent years how societal contributions made by university research can be identified, monitored and assessed (Muhonen, Benneworth, & Olmos-Penuela, 2020). Interviews, focus groups and other qualitative methods (e.g. document analysis), are increasingly being used to investigate productive interactions between researchers and societal stakeholders to identify various impact pathways through which science contributes to society (Boshoff & Sefatsa, 2019; Molas-Gallart & Tang, 2011; Spaapen & van Drooge, 2011). This line of inquiry reveals a general distinction between direct, indirect and financial forms of "productive interactions" (Spaapen & van Drooge, 2011) – facilitated by diverse and contingent impact pathways. Contributions of university research to society at large hinge on complex interplays between scholars, the knowledge they produce and the socio-cultural context they are situated in (Muhonen et al., 2020) Its study is primarily aimed at demonstrating which strategies researchers can adopt to achieve SIUR (de Jong, Barker, Cox, Sveinsdottir, & Van den Besselaar, 2014).This body of work is therefore less concerned with exploring how SIUR can be integrated in existing evaluation models, and more about demonstrating how it can be successfully pursued (Boshoff & Sefatsa, 2019).

Although the complicated and interactive nature of SIUR is now broadly recognized (Muhonen et al., 2020) and the view that it generally consists of iterative processes of contribution is uncontroversial (Sivertsen & Meijer, 2020), some studies remain committed to demonstrate how SIUR can be causally attributed to some extent (Noyons, 2019). Based on the recognition that a multiplicity of societal impact forms arise from the broad dissemination and popular uptake of knowledge created by university research, metric approaches have been applied to alternative circuits of communication (Bornmann, 2014). Generally, these altmetrics attribute a certain degree of SIUR to particular publications based on their diffusion on social media platforms, mostly Twitter (Robinson-Garcia et al., 2018). Here, the argument is that tracking the circulation of research outcomes among social media users offers an indication of societal interest and appreciation, which differs considerably from SIUR assessments made by peer reviewers (Bornmann, Haunschild, & Adams, 2019). But due to the affordances of social media platforms, even its advocates caution against simplistic impact attributions using altmetrics (Noyons, 2019). For example, it is fairly easy to boost altmetrics scores artificially through "academic spamming" (Erdt, Nagarajan, Sin, & Theng, 2016). At the same time, social media are also segmented into online communities, and in some disciplines scholarly work tends to circulate primarily among academic users (Zhou & Na, 2019), again dissuading all too simplistic attributions of SIUR using altmetrics.

Calibrated attributions: Written press articles as alternative objects for SIUR assessment

While attribution-based approaches towards SIUR continue to offer only partial indications of societal impact, it is at this time not warranted to disregard them altogether. It is certainly productive to emphasize the longitudinal, iterative and contingent nature of SIUR (Muhonen et al., 2020), but in some cases the possibility remains of causally attributing certain forms of societal impact (Sivertsen & Meijer, 2020). Indeed, many research activities address issues of direct societal relevance, and are carried out with the express intention to achieve some form of demonstrable impact, of which projects funded to meet "grand societal challenges" or the UN Sustainable Development Goals are evident examples (Holbrook & Frodeman, 2011). When researchers predetermine particular activities to accomplish societal impact, these can offer a modality to assess SIUR. Accordingly, it is opportune to explore how such instances might be integrated in quality evaluations and scientific excellence assessments. In the first place, this hinges on categorizing particular ambitions associated with the achievement of societal impact by university research. Secondly, it requires the identification of objects that can substantiate such accomplishments and allow for (partial) attribution of SIUR. Impact narratives like those required by the REF (Chowdhury et al., 2016; Hill, 2016) and multimethodological inquiry into productive interactions and impact pathways (Boshoff & Sefatsa, 2019; de Jong et al., 2014) are less concerned with demonstrating linear, causal results, but hint at particular measures taken by researchers to accomplish impact. Many of these do not lend themselves to systematic measurement, but others do provide opportunities for partial, less labour-intensive forms of SIUR assessment. Of specific interest here is the potential some see in (non-academic) media to pursue SIUR, particularly for humanities and social sciences (Muhonen et al., 2020). Inquiry into social and cultural phenomena may invite some to disseminate ideas an observations not only through formal circuits of communication of peer-reviewed scientific publications, but also to articulate them in the wider public sphere (Benneworth, Gulbrandsen, & Hazelkorn, 2016). Of course, the adoption of outreach strategies and science communication is not restricted to the social sciences and humanities; most scientific areas can plausibly effectuate some form of SIUR by communicating to larger audiences (Kassab, 2019). Accordingly, in qualitative assessments like the REF – which requires self-reported impact narratives enriched with documentation (Watermeyer & Chubb, 2019) – participation in public debate by pursuing mainstream media presence is both proposed and accepted as indicative of SIUR (Kassab, 2019).

At a glance, one could surmise that the documented presence of research or researchers in the written press provides a suitable object to systematically attribute (a degree of) SIUR by. In fact, doing so could compare favourably to the circulation on social media platforms currently prioritized by altmetrics – given that academic spamming is less likely in linear outlets, where its expected readership is more diverse (Erdt et al., 2016; Zhou & Na, 2019). Yet various complexities do arise when reflecting on the relation between popular media and SIUR that require consideration and mediation. Naturally, newspaper articles are largely "citable", and in many cases offer a disambiguated object of analysis and measurement (Didegah, Bowman, & Holmberg, 2018). However, their substance – and logically their presentation of research and researchers – varies considerably, and it is simply inappropriate to accept every substantiated instance of written press attention or intervention as an indication of SIUR. Indeed, in attributing some form of societal impact based on the analysis of written press documents, care must be taken not only to avoid undue accreditation, because researchers or their work are not necessarily included for reasons that appeal to a conceptualisation of SIUR. Equal care must go to circumventing undesirable behavioural changes, of which the subjection of the research agenda to considerations of popular appeal is but one example. Before written press coverage can be seriously considered as a partial indicator in systematic SIUR assessment, the presentation of research and researchers must be explored to further calibrate its use as an evaluation object.

Scholars in broadsheets: A disciplinary case on Flemish sociologists in the written press

A case study into written press coverage on Flemish sociologists offers an initial, descriptive exploration of these complexities. Here, the focus on sociology reflects the more pressing demand for alternative quality indicators in social sciences and humanities disciplines (Gijselinckx & Steenssens, 2011), but it goes without saying that the future operationalisation of written press articles in SIUR assessment must be predicated on multifocal analysis of its benefits and pitfalls. The study inquired when and how 360 faculty members (from researchers to full professors), recently employed at one of the four Flemish sociology departments were present in domestic written media with national circulation. This was effectuated by querying their full names and the term "onderzoek" (research) in the 2010-2020 period in Gopress – an online database operated by press agency Belga that collects all articles published by outlets with national circulation (e.g. newspapers, broadcaster websites, magazines). In doing so, it differs from existing studies on scholarly attention in the written press, which have mostly employed self-reporting (e.g. Bauer and Jensen (2011), Dudo (2012)). The present study, by contrast, includes not only coverage of which researchers are aware, but also written press documents in which they are cited, mentioned or referred to without their knowledge – thus providing a more in-depth perspective crucial to SIUR assessment.

When looking at observable differences in written press coverage (#1656 articles) of sociology researchers ($N = 360$) in the 2010-2020 period, data show that less than half of all sociologists (40.3%) were present at least once in mainstream reporting, with an average of 9.2 ($SD = 38.9$) and a median value of 0 mentions. When observing for a recent one-year period (2019-2020) the percentage is less than a quarter (23.3%), with an average of 1.9 ($SD = 9.5$) and a median value of 0 mentions. These averages are misleading however, as there are substantial outliers present: the distribution of media mentions is skewed as the bulk of the mentions come from a small number of outliers. In addition, these numbers must be nuanced due to the presence of early career researchers and professorial staff in the same data set, and due to double entries because of media concentration in the domestic press landscape (Hendrickx & Ranaivoson, 2019).

After correcting for outliers, predoctoral researchers found themselves 1.1 ($SD = 0.3$) times on average in the written press, mid-level (i.e. junior & senior) researchers did so 4.9 ($SD = 1.2$) and 4.5 ($SD = 1.6$) times respectively, and those with a professorial position boasted 15.9 ($SD = 2.9$) instances on average. Researcher status (measured by #publications and career position standardized) was found to correlate positively ($\beta = .45$, $p < .001$) with press attention (#mentions in 2010-2020 period) – suggesting that higher status researchers have easier access to popular reporting.

Although the foregoing cautions against all too simplistic conclusions about the number of instances of researchers appearing in the written press, it does not address whether it is in the first place legitimate or desirable to attribute a degree of SIUR to such instances.

Further analysis of the collected documents revealed that more than half of the articles (54.3%) were short (less than 250 words) and declarative in nature – offering factual coverage of conducted studies and their results. Even though this type of presence in the written press could effectuate some form of SIUR (Morton, 2015), it is mostly descriptive in nature, and is in the first place committed to informing the general public about sociological research and its results.

A second type of presence is found in so called "think pieces" (in-depth discussion of one theme) that make up almost a quarter of the sample (23%), and is characterized by active operationalisations of research – either in the form of direct quotes of researchers or their publications, or through paraphrases and references. These longer texts featured in-depth engagements with sociological research – often including the views of various scholars – to substantially address a particular societal issue. These subjects tended to reflect the core disciplinary expertise of sociology, as socio-economical topics, discrimination, politics, integration and education were the dominant topics respectively. Opinion pieces, finally, comprised a third important category, with 12.1% of samples. In contrast to other types, these tended to be authored by researchers themselves, and often constituted a public intervention based on their particular scholarly expertise. While these were not necessarily a research outcome in the strict sense, they tended to apply common sociological analytics to applied societal phenomena.

These exploratory observations point to a distinction between coverage of, operationalizations of, and interventions by research in the public debate. Based on this taxonomy, an extension of altmetrics to press coverage appears to have potential, but operationalizations must differentiate between popular attention for research and its outcomes and active research-based contributions to public debates. Accordingly, further research is needed to develop tools that efficiently differentiate between written press documents that demonstrate the popularity of research – and successful science communication – and public operationalizations of research that indicate a more profound societal impact.

References

Agrawal, A., & Henderson, R. (2002). Putting patents in context: Exploring knowledge transfer from MIT. *Management Science, 48*(1), 44-60.

Bauer, M. W., & Jensen, P. (2011). The mobilization of scientists for public engagement. *Public Understanding of Science, 20*(1), 3-11.

Benneworth, P., Gulbrandsen, M., & Hazelkorn, E. (2016). *The Impact and Future of Arts and Humanities Research*. London: Palgrave Macmillan Ltd.

Benneworth, P., & Olmos-Peñuela, J. (2018). Reflecting on the tensions of research utilization: Understanding the coupling of academic and user knowledge. *Science and Public Policy, 45*(6), 764-774.

Bornmann, L. (2013). What Is Societal Impact of Research and How Can It Be Assessed? A Literature Survey. *Journal of the American Society for Information Science and Technology, 64*(2), 217-233.

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics, 8*(4), 895-903.

Bornmann, L., Haunschild, R., & Adams, J. (2019). Do altmetrics assess societal impact in a comparable way to case studies? An empirical test of the convergent validity of altmetrics based on data from the UK research excellence framework (REF). *Journal of Informetrics, 13*(1), 325-340.

Boshoff, N., & Sefatsa, M. (2019). Creating research impact through the productive interactions of an individual: an example from South African research on maritime piracy. *Research Evaluation, 28*(2), 145-157.

Chowdhury, G., Koya, K., & Philipson, P. (2016). Measuring the Impact of Research: Lessons from the UK's Research Excellence Framework 2014. *PLoS One, 11*(6), e0156978.

Czarnitzki, D., Rammer, C., & Toole, A. A. (2014). University spin-offs and the "performance premium". *Small Business Economics, 43*(2), 309-326.

de Jong, S., Barker, K., Cox, D., Sveinsdottir, T., & Van den Besselaar, P. (2014). Understanding societal impact through productive interactions: ICT research as a case. *Research Evaluation, 23*(2), 89-102.

Didegah, F., Bowman, T. D., & Holmberg, K. (2018). On the Differences Between Citations and Altmetrics: An Investigation of Factors Driving Altmetrics Versus Citations for Finnish Articles. *Journal of the Association for Information Science and Technology, 69*(6), 832-843.

Donovan, C. (2011). State of the art in assessing research impact: introduction to a special issue. *Research Evaluation, 20*(3), 175-179.

Donovan, C. (2019). For ethical 'impactology'. *Journal of Responsible Innovation, 6*(1), 78-83.

Dudo, A. (2012). Toward a Model of Scientists' Public Communication Activity. *Science Communication, 35*(4), 476-501.

Erdt, M., Nagarajan, A., Sin, S. C., & Theng, Y. L. (2016). Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media. *Scientometrics, 109*(2), 1117-1166.

Fochler, M., Felt, U., & Muller, R. (2016). Unsustainable Growth, Hyper-Competition, and Worth in Life Science Research: Narrowing Evaluative Repertoires in Doctoral and Postdoctoral Scientists' Work and Lives. *Minerva, 54*(2), 175-200.

Gijselinckx, C., & Steenssens, K. (2011). *Naar waarde geschat: Valorisatie van onderzoek in de humane en sociale wetenschappen*. Brussel: Vlaamse Raad voor Wetenschap en Innovatie.

Hendrickx, J., & Ranaivoson, H. (2019). Why and how higher media concentration equals lower news diversity - The Mediahuis case. *Journalism*, 1464884919894138.

Heyeres, M., Tsey, K., Yang, Y., Yan, L., & Jiang, H. (2019). The characteristics and reporting quality of research impact case studies: A systematic review. *Eval Program Plann, 73*, 10-23.

Hill, S. (2016). Assessing (for) impact: future assessment of the societal impact of research. *Palgrave Communications, 2*(1), 16073.

Holbrook, J. B., & Frodeman, R. (2011). Peer review and the ex ante assessment of societal impacts. *Research Evaluation, 20*(3), 239-246.

Kassab, O. (2019). Does public outreach impede research performance? Exploring the 'researcher's dilemma' in a sustainability research center. *Science and Public Policy, 46*(5), 710-720.

Matt, M., Gaunand, A., Joly, P. B., & Colinet, L. (2017). Opening the black box of impact – Ideal-type impact pathways in a public agricultural research organization. *Research Policy, 46*(1), 207-218.

Molas-Gallart, J., & Tang, P. (2011). Tracing 'productive interactions' to identify social impacts: an example from the social sciences. *Research Evaluation, 20*(3), 219-226.

Muhonen, R., Benneworth, P., & Olmos-Penuela, J. (2020). From productive interactions to impact pathways: Understanding the key dimensions in developing SSH research societal impact. *Research Evaluation, 29*(1), 34-47.

Noyons, E. (2019). Measuring societal impact is as complex as ABC. *Information Science, 4*(3), 6-21.

Rijcke, S. d., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—a literature review. *Research Evaluation, 25*(2), 161-169.

Robinson-Garcia, N., van Leeuwen, T. N., & Rafols, I. (2018). Using altmetrics for contextualised mapping of societal impact: From hits to networks. *Science and Public Policy, 45*(6), 815-826.

Sivertsen, G., & Meijer, I. (2020). Normal versus extraordinary societal impact: how to understand, evaluate, and improve research activities in their relations to society? *Research Evaluation, 29*(1), 66-70.

Spaapen, J., & van Drooge, L. (2011). Introducing 'productive interactions' in social impact assessment. *Research Evaluation, 20*(3), 211-218.

Watermeyer, R., & Chubb, J. (2019). Evaluating 'impact' in the UK's Research Excellence Framework (REF): liminality, looseness and new modalities of scholarly distinction. *Studies in Higher Education, 44*(9), 1554-1566.

Watermeyer, R., & Hedgecoe, A. (2019). Selling 'impact': peer reviewer projections of what is needed and what counts in REF impact case studies. A retrospective analysis. *Journal of Education Policy, 31*(5), 651-665.

Zhou, Y. F., & Na, J. C. (2019). A comparative analysis of Twitter users who Tweeted on psychology and political science journal articles. *Online Information Review, 43*(7), 1188-1208.

Additional Context Helps! Leveraging Cited Paper Information to Improve Citation Classification

Kamal Kaushik Varanasi[1], Tirthankar Ghosal[2], and Valia Kordoni[3]

[1] kamalkaushikv@gmail.com, [2] tirthankar.pcs16@iitp.ac.in
Indian Institute of Technology Patna, Bihar (India)
[3] evangelia.kordoni@anglistik.hu-berlin.de
Humboldt Universität zu Berlin (Germany)

Abstract

With the rapid growth in research publications, automated solutions to tackle scholarly information overload is growing more relevant. Correctly identifying the intent of the citations is one such task that finds applications ranging from predicting scholarly impact, finding idea propagation, to text summarization to establishing more informative citation indexers. In this in-progress work, we leverage the cited paper's information and demonstrate that this helps in the effective classification of citation intents. We propose a neural multi-task learning framework that harnesses the structural information of the research papers and the relation between the citation context and the cited paper for citation classification. Our initial experiments on three benchmark citation classification datasets show that with incorporating cited paper information (title), our neural model achieves a new state of the art on the ACL-ARC dataset with an absolute increase of **5.3%** in the F1 score over the previous best model. Our approach also outperforms the submissions made in the 3C Shared task: Citation Context Classification with an increase of **8%** and **3.6%** over the previous best Public F1-macro and Private F1-macro scores respectively.

Introduction

Citations are crucial in analyzing scientific works and for understanding the link between different research articles. They act as trackers of the direction of research in a field and as an important measure in understanding the impact of research articles, venues, researchers, etc. Citations may also have different nature. Authors may cite a research publication in different ways. For example - a citation might indicate motivation or usage of a method from a previous work or a comparison of results of various works. So, identification of the intent behind a citation is crucial for automated analysis of academic literature. Most of the research works in the field of citation classification provide too fine-grained citation categories, example-(Stevens and Giuliano (1965); Moravcsik and Murugesan (1975)), so only a handful of these are used for automated analysis of the scientific publications. To overcome these problems, Jurgens et al. (2018) proposed a six category classification scheme. Then, Cohan et al. (2019) used a different scheme that had only three classification categories. More recently, Pride et al. (2020) proposed a classification scheme similar to Jurgens et al. (2018).

Table 1. Examples of citations with cited paper titles and intents.

Citation context	*Cited Paper Title*	*True Label*
She evaluates 3,000 German verbs with a token frequency between 10 and 2,000 against the Duden (@@CITATION).	duden—das stilworterbuch duden—the style dictionary	BACKGROUND

Jurgens et al. (2018) used a set of engineered features like (1) Pattern based features (2) Topic based features (3) Prototypical Argument features for this task. While recently, Cohan et al. (2019) argued that features based on the structural properties related to scientific literature are more effective than the predefined hand engineered domain-dependent features or external resources. We argue that in addition to leveraging the structural information related to the scientific discourse, utilizing the cited paper information as additional context can significantly improve the performance. In the example from table 1, it is evident that the instance seems less ambiguous and easier to classify after accessing the cited paper title information in addition to the citation context. To tackle these problems, we propose a Multi Task Learning framework that incorporates three scaffolds, including a cited paper title scaffold that leverages the relationship between the citation context and the cited paper title. The other two scaffolds are the structural scaffolds to leverage the relationship between the structure of the research papers and the intent of the citations. These two scaffolds are inspired by the work done in Cohan et al. (2019). We explain these scaffolds in detail in Table 4 under the Model Section.

Dataset Description

Table 2 shows the classification categories of different datasets and Table 3 shows the corresponding data statistics. Please note that we retrieve the cited paper title scaffold data from the target datasets and the SciCite dataset includes the data corresponding to the structural scaffolds.

Table 2.Intent categories of different datasets.

Dataset	*Citation Intent Categories*
SciCite	BACKGROUND, METHOD, RESULT_COMPARISON
ACL-ARC/3C Challenge dataset	BACKGROUND, USES, COMPARE_CONTRAST, MOTIVATION, EXTENSION, FUTURE.
Section title scaffold data (91412 instances)	INTRODUCTION, CONCLUSION, EXPERIMENTS, METHOD, RELATED WORK
Citation worthiness scaffold data (73484 instances)	TRUE, FALSE

Table 3.Cross-study comparison of different datasets.

Dataset	*Papers*	*Annotated by*	*Citations*	*Intent categories*	*Discipline(s)*
SciCite	6,627	Volunteers	11,020	3	Comp. Sci/Medicine
3C Challenge	883	Paper authors	3,000	6	Multi-disciplinary
ACL-ARC	185	Domain Experts	1,989	6	Comp. Science

Model

We propose a Multitask Learning (Caruana, 1997) model with the main task of citation intent classification along with a total of three auxiliary tasks (scaffolds). The knowledge acquired by the auxiliary tasks helps the model to learn optimal parameters for the main task.

Table 4. Scaffolds in our Multi-tasking Approach

Scaffolds	*Description*
Section Title	This task is related to predicting the section under which the citation occurs, given a citation context. In general, researchers follow a standard order while presenting their scientific work in the form of sections. Citations may have different nature according to the section under which they are cited. Hence, the intent of the citation and the section are related to each other. For example, the results-comparison related citations are often cited under the Results section.
Citation Worthiness	This task is related to predicting whether a sentence needs a citation or not, i.e., it is the task of classifying whether a sentence is a citation text or not.
Cited Paper Title	Sometimes a citation context might be ambiguous, making it difficult to predict the intent of the citation correctly. In such cases, information from the cited paper like the abstract of the paper, title of the paper, etc. may provide some additional context that can assist in identifying the appropriate intent behind that citation. This auxiliary task helps the model to learn these nuances by leveraging the relationship between the citation context and the cited paper. We use a concatenated vector of citation context and the cited paper title fields from the target dataset as the input for this task.

We use these auxiliary tasks only while training/fine-tuning the model. Our model architecture is shown in Figure 1.

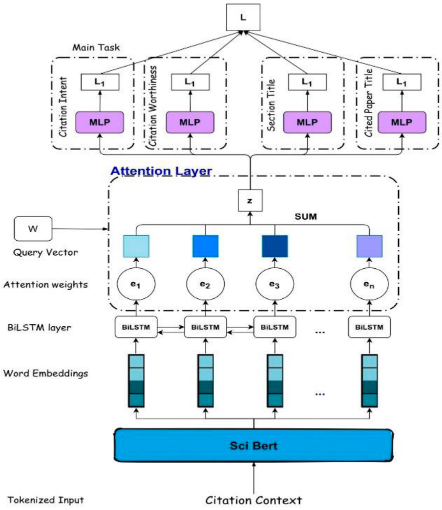


Figure 1. The architecture of our proposed model. The main task MLP is for prediction of citation intents (top left) followed by three MLPs for section title, citation worthiness, and cited paper title scaffolds

Model Structure

Let C be the tokenized citation context of size n. We pass it onto the SciBERT (Beltagy, 2019) model with pre-trained weights to get the word embeddings of size (n,d_1) i.e. we have the output as $x = \{ x_1, x_2, x_3, \ldots\ldots x_n \}$ where $x_i \in R^{d1}$. Then we use a Bidirectional long short-term memory (BiLSTM) network with a hidden size d_2 to get an output vector h of size ($n,2d_2$). We pass h to the dot-product attention layer with query vector w to get an output vector z which represents the whole input sequence.

$$h_i = [LSTM(x,i); LSTM(x,i)] \quad (1)$$

$$\alpha_i = softmax(w^T h_i) \quad (2)$$

Here, α_i represents the attention weights.

$$z = \sum_{i=1}^{n} \alpha_i h_i \quad (3)$$

Now, we pass the attention representation vector z to m MLPs related to the m tasks with $Task_1$ as the main task and $Task_i$ as the m-1 scaffold tasks, where $i \in [2, m]$, to get an output vector $y = \{ y_1, y_2, y_3, \ldots\ldots y_m \}$:

$$y_i = softmax(MLP_i(z)) \quad (4)$$

For each task, we use a Multi Layer Perceptron (MLP) followed by a softmax layer to obtain the class with the highest class probability. The parameters of a task's MLP are the specific parameters of that task and the parameters in the lower layers (parameters till the attention layer) are the shared parameters.

Training

We train our model only on the SciCite dataset. Then we fine-tune on the target dataset (ACL-ARC or 3C Challenge). We use the pre-trained scibert scivocab uncased model trained on a corpus of 1.14M papers and 3.1B tokens to get the 768-dimensional Word Embeddings. While training on the SciCite dataset, we only train the two structural scaffolds which are - 1. Citation Worthiness scaffold, 2. Section Title scaffold, along with the main task. While fine-tuning on the target datasets, we use the Cited paper title scaffold only, while freezing the task specific parameters of the other two scaffolds, learned during the training on the SciCite dataset. We compute the loss function as:

$$L = \sum_{(x,y) \in D_1} L_1(x,y) + \sum_{i=2}^{n} \lambda_i \sum_{(x,y) \in D_i} L_i(x,y) \quad (5)$$

where D_i is the labeled dataset corresponding to $task_i$, λ_i is the hyperparameter that specifies the sensitivity of the model to each specific task, L_i is the loss corresponding to $task_i$. In each training epoch, we take a batch with equal number of instances from all the tasks and calculate the loss as specified in Equation 5, where $L_i = 0$ for all the instances of other tasks, $task_k$ where $k \neq i$. Then, we perform back propagation and update the parameters using the AdaDelta optimizer with gradient clipping.

Experiments

Baselines

We have worked on multiple baseline models to compare their performance on the ACL-ARC and the 3C Challenge datasets.

Table 5. Baselines and Proposed System.

Baselines	*Description*
BiLSTM+Attention (with SciBERT)	This baseline has a similar structure as our proposed model until the attention layer. It only has one MLP related to the main task and optimizes the network for the main loss.
3C Shared Task Submission 1	This system submission has achieved the best submission results on the 3C Challenge dataset in the Kaggle 3C Shared Task. The model is a Passive Aggressive Classifier with a concatenated vector (including the citing paper title, cited paper title and the citation context) as the input.
Cohan et al. (2019)	The model has reported state-of-the-art results on the ACL-ARC dataset. It incorporates a multi task learning framework with two structural scaffolds predicting the section title and citation worthiness, given the citation context.
Representation Model	The model framework for this baseline incorporates the concatenation of two representation vectors which is passed on to a MLP for classification. We get the first representation from the attention layer of the pretrained first baseline by passing citation context and the cited title as input. We use the pre-trained Cohan et al. (2019) model, trained on SciCite to get the predicted labels on the target dataset. Then, we infuse this external knowledge with the citation context and pass it to the first baseline to obtain the second attention layer representation.
Late Fusion Model	This baseline model has a similar structure to that of the first baseline. We use the pre-trained Cohan et al. (2019) model, trained on SciCite to get the citation intent, section title and the citation worthiness labels. We concatenate these labels with the output of the attention layer of this baseline and pass it to a MLP for prediction.

Table 6. Results on the ACL-ARC and the 3C Challenge datasets. The first two columns (Macro F1 score and Accuracy) correspond to the results on the ACL-ARC dataset. The last two columns (Public and Private F1 scores) are the results on the 3C Challenge dataset.

Category	*ACL-ARC*		*3C Challenge dataset*	
	Macro F1 Score	*Accuracy*	*Public F1*	*Private F1*
BiLSTM + Attention (with SciBERT)	57.1	63.3	27.8	23.9
Cohan et al. (2019)	67.9	76.2	22.4	**25.2**
Kaggle 3C Shared Task Submission 1	-	-	21.5	20.6
Our Model	**73.2**	**77.0**	**29.5**	24.2
Representation Model	38.2	54.7	20.6	23.1
Late Fusion Model	48.3	61.9	22.4	22.4

Results

Our results for the ACL-ARC and the 3C challenge datasets are shown in Table 6. We observe that as compared to the first baseline "BiLSTM+Attention (with SciBERT)", Cohan et al. (2019) achieves an F1 macro score of 67.9 (Δ = 10.8) and a validation accuracy of 76.2 (Δ =

12.9) on the ACL-ARC dataset indicating the fact that leveraging the structural information of a research work helps the model to learn more effectively. Out of all the other baselines, our model achieves the best results with an F1 score of 73.2, a significant improvement over the previous state of the art results of Cohan et al. (2019) (Δ = 5.3) and a validation accuracy of 77.0 (Δ = 0.8) on the ACL-ARC dataset. This clearly demonstrates the efficacy of using the three scaffolds in a transfer learning framework for this task. For the last two baselines that are mainly based on fusing external knowledge obtained by using the pre trained Cohan et al. (2019) model, we find a significant dip in the performance. This suggests that this external knowledge does not provide any useful signals beyond what the first baseline already learns from the data. The results on the 3C Challenge dataset also show similar patterns.

Analysis

To gain more insight into how the scaffolds are helping the model, we consider examples from the ACL-ARC and the 3C Challenge datasets and compare the predictions of the simple baseline *'BiLSTM+Attention (with SciBERT)'*, the previous state of the art *'Cohan et al. (2019)'*, and our best-proposed model *'BiLSTM+Attention (with SciBERT)+three scaffolds'*. In table 7, the first example is from the ACL-ARC dataset with true label *COMPARE*, the simple baseline and the Cohan et al. (2019) incorrectly predict it as *MOTIVATION* and *BACKGROUND* respectively, whereas our model predicts it correctly. The simple baseline does not include any scaffold, while the Cohan et al. (2019) and our model incorporate a multi task learning framework. Note that our model is similar to that of Cohan et al. (2019) but includes three scaffolds (two structural scaffolds + cited paper title scaffold). The second example is from the 3C Challenge dataset where the true label is *BACKGROUND*, the Cohan et al. (2019) model is probably distracted by the phrase "use", so it classifies it incorrectly as *USE*, whereas our model correctly classifies it. Note that our model also consists of additional information from the cited paper (title) which provides additional context, thus helping it to classify better.

Table 7. A sample of predictions of the models on examples from the ACL-ARC and the 3C Challenge datasets.

Example	*Model*	*Predicted Label*	*True Label*
The advantage of tuning similarity to the application of interest has been shown previously by CITATION.	BiLSTM + Attention (with SciBERT)	MOTIVATION	COMPARE
	Cohan et al.	BACKGROUND	COMPARE
	Our Model	COMPARE	COMPARE
Others use concepts such as expansion and contraction (Mattsson, 1987); extension and consolidation CITATION and splitting and joining (Hertz, 1996)	Cohan et al.	USE	BACKGROUND
	Our Model	BACKGROUND	BACKGROUND
We experiment with four learners commonly employed in language learning: Decision List (DL): We use the DL learner as described in CITATION, motivated by its success in the related tasks of word sense disambiguation (Yarowsky , 1995) and NE classification (Collins and Singer , 1999) .	Our Model	USE	MOTIVATION

Error Analysis

We investigate the type of errors made by our proposed model on the two datasets. We found it surprising to note that in case of the ACL-ARC dataset, the model has more tendency to make false positive errors in the *COMPARE* category, although it being the second most dominating category. Whereas in the case of the 3C Challenge dataset, it makes many false positive errors in the *BACKGROUND* category. To overcome this problem of overfitting, we decided to use some oversampling techniques like SMOTE, but we still did not get any significant improvements. Figure 2 shows the confusion matrix of our best model on the two datasets. We also found out that some errors are due to ambiguity in the citation context as well as the title of the cited paper. We can avoid them by providing some additional context apart from the cited paper title information (for example, providing abstract from the cited paper, etc). In the last example from Table 7, the model is probably distracted by the phrases "We use" and "as described in CITATION", leading to an inference that there is a usage of a method from the cited paper, instead of considering the latter part of the sentence that describes the motivation. This is likely due to the small number of training instances in the *MOTIVATION* category, preventing the model from learning such subtle details.

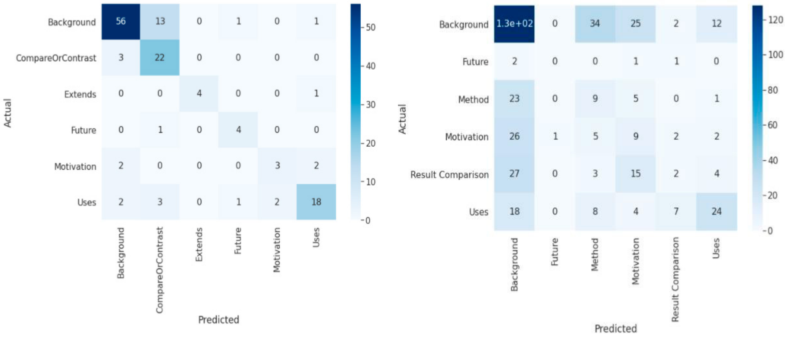


Figure 2. Confusion matrix showing the classification errors of our best model on the ACL-ARC (test size: 139) and the 3C Challenge datasets (test size: 400) respectively.

Conclusion and Future Work

In this work, we demonstrate that the structural information related to a research paper and additional context (title information) of the cited paper can be leveraged to effectively classify the intent of the citations. A future line of research would be to use the abstract of the cited paper as further contextual information for the task and also to investigate alternative approaches for solving the issue of overfitting on the 3C Challenge dataset.

References

Beltagy, I., Lo, K. & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *EMNLP*.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28, 41–75.

Cohan, A., Ammar, W., Zuylen, M. & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers) (*NAACL*) (2019), 3586–3596.

Jurgens, D., Kumar, S., Hoover R., McFarland D. & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *TACL*, 6, 391-406.

Moravcsik, M. & Murugesan, P. (1975). Some results on the function and quality of citations. *Social studies of science*, 5(1), 86–92.

Pride, D. & Knoth, P. (2020). An Authoritative Approach to Citation Classification. In: *ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20)*, 1-5 Aug 2020, Virtual China.

Stevens, M. & Giuliano, V. (1965). *Statistical Association Methods for Mechanized Documentation: Symposium Proceedings, Washington, 1964*, volume 269. US Government Printing Office.

Thematic influence on academic impact according to types of collaboration: an analysis of the library and information science field from 2015 to 2019

A. Velez-Estevez[1], P. Garcia-Sanchez[2], J.A. Moral-Munoz[3] and M.J. Cobo[1]

[1] *{antonio.velez, manueljesus.cobo}@uca.es*
Department of Computer Science and Engineering, University of Cadiz, Puerto Real (Spain)

[2] *pablogarcia@ugr.es*
Department of Computer Architecture and Technology, University of Granada, Granada (Spain)

[3] *joseantonio.moral@uca.es*
Department of Nursing and Physiotherapy, University of Cadiz, Cadiz (Spain)
Institute of Research and Innovation in Biomedical Sciences of the Province of Cadiz (INiBICA), Cadiz (Spain)

Abstract

The international collaboration has become usual in the research activity. Moreover, international collaboration tends to lead papers to have higher citations than papers with national collaboration. This fact has been previously studied finding a difference of impact in several areas. However, some reasons that could influence this difference of impact have not been studied for the Library and Information Science (LIS). Thus, in this contribution we analyze 22,127 papers from the years 2015 to 2019. These papers were classified under three classes: local, national, or international collaboration. They were analyzed separately by means of science mapping analysis and performance measures. The results show that there is a difference in impact as well as thematical differences among the collaboration types.

Introduction

In the last years, the increase of complexity of the science, and the need of being competitive, has led countries to cooperate with other countries. Therefore, the research policies implemented by countries have encouraged the mobility of researchers as well as the funding of international projects (Sugimoto et al., 2017; Suresh, 2012). Consequently, the researchers need to collaborate with foreign colleagues in order to achieve a higher scientific impact (Chinchilla-Rodríguez, Sugimoto, & Larivière, 2019; Larivière, Gingras, Sugimoto, & Tsou, 2015). According to Adams (2013), that new trend in collaboration is known as the fourth age of research, which is driven by "international collaborations between elite research groups".

Due to the interest growth in assessing research, to implement new research policies, new tools dealing with the overwhelming amount of research output have been created to understand the science development or to detect patterns on it (Bornmann & Mutz, 2015; Fortunato et al., 2018; Milojević, 2015). Therefore, bibliographical networks (Batagelj & Cerinšek, 2013) and science mapping analysis tools help scientists of science on this task (Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2012; Moral-Muñoz, Herrera-Viedma, Santisteban-Espejo, & Cobo, 2020).

There are several studies analyzing the consequences that international collaboration may have in academic impact. Indeed, the difference between the impact of academic collaboration has been studied previously in several areas, and, in general, international collaboration entails a higher impact rather than other types of collaboration (Gazni, Sugimoto, & Didegah, 2012; Persson, 2010). It has also been studied in particular areas (Khor & Yu, 2016; Polyakov, Polyakov, & Iftekhar, 2017; Rousseau & Ding, 2016; Sooryamoorthy, 2017) as well as in Library and Information Science (LIS) (Asubiaro, 2019; Sin, 2011; Sooryamoorthy, 2017).

Nonetheless, other influences that might affect the impact of academic research have been studied, such as publishing in open access (Gabrielle Breugelmans et al., 2018), the leadership impact (Chinchilla-Rodríguez et al., 2019) or if this increase in impact occurs in all the research fields (Polyakov et al., 2017). Moreover, it has been studied that government funding is statistically nonsignificant in international collaboration in the OCDE countries (Leydesdorff, Bornmann, & Wagner, 2019). Furthermore, the thematical landscape has been studied previously in the LIS category using co-word and co-citation analysis (Galvez, 2018; Hu, Hu, Deng, & Liu, 2013; Olmeda-Gómez, Ovalle-Perandones, & Perianes-Rodríguez, 2017). However, the themes treated in the academic collaboration and its possible influence in the impact difference of academic collaboration on LIS field have not been studied.

Thus, the goal of this contribution is to analyze the thematical differences or similarities that might affect the academic impact differences between types of collaboration that other studies have confirmed in the specific ambit of the LIS category. To do so, the papers of this category published during the period 2015-2019 have been downloaded from Web of Science and analyzed by means of science mapping analysis and performance analysis. The main hypothesis of this contribution states that one of the reasons for the difference in academic impact of the Library & Information Science category are the themes treated.

The rest of the contribution is organized as follows. First, the methodology used to carry out the analysis of the metadata is explained. Then, the results obtained are presented and described. Next, the discussion regarding all the results from a global perspective and the findings of the study are shown. Finally, the conclusions and future work are presented.

Methodology

This section describes the methodology followed to carry out the analysis. In order to find out the differences or similarities between the types of collaboration, it is necessary to define them. According to previous studies (Chinchilla-Rodríguez et al., 2019; Gazni et al., 2012), there are three types of collaboration: local (intra-institutional collaboration), national (collaboration with colleagues from same country) and international collaboration (collaboration with colleagues from foreign countries).

In order to perform any bibliometric analysis, the first step is to obtain the data to be analyzed. It can be done by retrieving the data from the several bibliographical databases available, such as Scopus, Web of Science, Dimensions or ScholarMetrics, among others. The analysis that will be performed needs the following metadata: author's keywords, author's affiliations, and publications year. Therefore, the bibliographical database must offer this kind of metadata.

For this study, the Web of Science database was selected to download the papers metadata. The query to perform is filtered by the year, the paper category, the database editions, and the document type. The limits were the following:

- Documents between 2015 and 2019, both included.
- Documents under the "Information Science & Library Science" category.
- Documents indexed on the Science Citation Index ™ Expanded (SCI) or Social Sciences Citation Index ® (SSCI).
- Articles or reviews.

Therefore, the advanced query to perform against the Web of Science database is WC="Information Science & Library Science" AND PY=2015-2019 AND DT= (ARTICLE OR REVIEW)"

To conduct the bibliometric analysis presented in this contribution, we have used the software SciMAT. One of the advantages of using SciMAT is that it is the only science mapping analysis tool with the de-duplication feature (Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2011a; Cobo et al., 2012; Moral-Muñoz et al., 2020). Moreover, it allows us to plot the detected themes in a strategic diagram which is a useful tool to uncover the differences among the different corpus and the differences of themes covered in each one.

The next step is to process the author's keywords to join the ones that represent the same concept. This step also involves the removal of terms with a broad meaning, known as stop words.

Once the author's keywords have been cleaned, the dataset is divided into the three types of collaboration mentioned above. It is done by computing the number of unique countries and organizations in the affiliation.

- The local (intra-institutional level) papers must satisfy the number of organizations and countries to be one.
- The national collaboration papers must satisfy the number of organizations to be more than one and to only have a single country.
- Finally, the international collaboration papers are classified under this tag when two or more countries are collaborating in them.

We should point out that mentioned sets are mutually exclusive.

After the split is performed, the analysis can be carried out. Firstly, the performance measures of each collaboration type are computed, and a comparison of academic impact in terms of citations is done. According to Thelwall (2016), citations should be averaged using the geometric mean, since they follow a discretized lognormal distribution. Moreover, the hypothesis will be checked by means of science mapping analysis and overlapping of the keywords between collaboration types.

Next, for each collaboration type, a bibliographical co-words network using the author keywords of each document will be built. In this network, the nodes will be the words and the edges will represent if those two keywords appear together in the papers or not. The network has two attributes: the frequency of a node, namely how many times such keyword appears in the documents; and the second, which is the co-occurrence frequency of two nodes (or the weight of an edge), which means the number of times that the two keywords appear together in a document. Next, a normalization process over the co-occurrence frequency is made by means of the equivalence index (Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2011b; Van Eck & Waltman, 2009). Once this step is done, clustering is performed applying the Leiden algorithm (Traag, Waltman, & van Eck, 2019). The Leiden clustering algorithm has been employed because it offers some guarantees such as well-connected communities, which is an advantage over other clustering algorithms (Traag et al., 2019).

Once the clusters (themes) are detected, some measures related to academic impact can be computed. It is done by performing the union of all the documents under the nodes of the clusters, and then, computing the measures over that set of documents. In this study, the number of documents, the citation average, the h-index and the mean normalized citation score (MNCS) (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011) will be computed for each theme detected in the three collaboration types. Moreover, these themes can be plotted out in a strategic diagram by computing the internal and external cohesion of the clusters, according to their density and centrality. Hence, the diagram classifies the themes in four categories: motor, basic and transversal, highly developed and isolated, and emerging or declining (Cobo et al., 2011b). It should be point out that a theme encapsulates a cluster. That is, a theme represents a set of strong connected keywords (sub co-word network), which received the name of the most central one.

In addition to this comparison, to give a performance perspective, the number of most cited keywords colliding among the three collaboration types will be measured and plotted out in a line chart. The number of shared keywords between the collaboration types will allow us to compare whether they are treating the same topics or not. The colliding keywords will be measured by the Jaccard index (Cobo et al., 2011b) for the *n* most cited keywords, with several values of *n*.

Results

The download was performed on October 6th, 2020 using the advanced query explained in the Methodology section, and it retrieved 22,127 records.

Performance analysis

The output of the performance analysis of the data by types of collaboration is shown in Table 1. The papers with no collaboration represent the half of the total, and national and international papers are the rest with similar amounts. A difference of impact can be noticed in the geometric mean of citations since it becomes higher when the collaboration level increases.

Table 1. Descriptive measures by collaboration type datasets

Metrics	*Local*	*National*	*International*
H-index	70	61	75
Citations (geometric mean)	4.14	5.13	6.3
Citations (standard deviation)	17.17	15.05	23.39
Citations (median)	3	4	5
%Papers	49.98%	23.39%	26.63%
Number of papers	10,158	4,755	5,413

We should point out that the sum of the papers of the three collaboration types is not equal to the total number of downloaded papers. This is due to 1,801 papers did not have the affiliation, i.e., neither the country nor the organization.

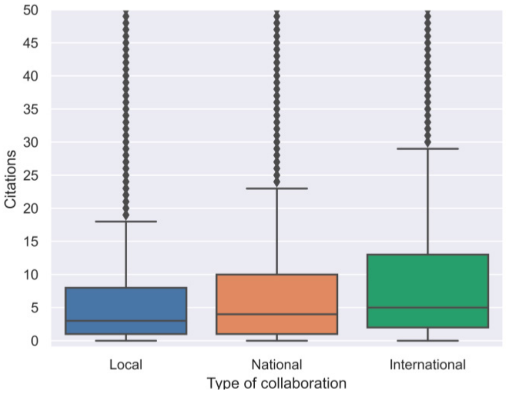


Figure 1. Boxplot diagram of local, national, and international collaboration citations.

In addition, as shown in Figure 1, the maximum of citations without outliers is higher when the collaboration level increases. Moreover, the outliers are more reduced in the national and international collaboration in comparison with the local collaboration type. The 25% of most cited papers increase when the academic collaboration is higher, this can be noticed by the 3rd quartile being higher in each of the collaboration types.

Themes in the collaboration types

A science mapping analysis has been made on each collaboration type to study the thematical differences and similarities between the collaboration types. The clusters detected for each collaboration type are shown in the strategic diagrams (Figures 2-4). For each of those clusters the number of documents, the citations average, the h-index and the MNCS have been computed. They are shown in Tables 2-4.

Prior to describing the results, is important to clarify that in the rest of the paper, the name of the themes will be in *italic*. Also, it is important to point out that concepts related to themes are nodes under the detected clusters.

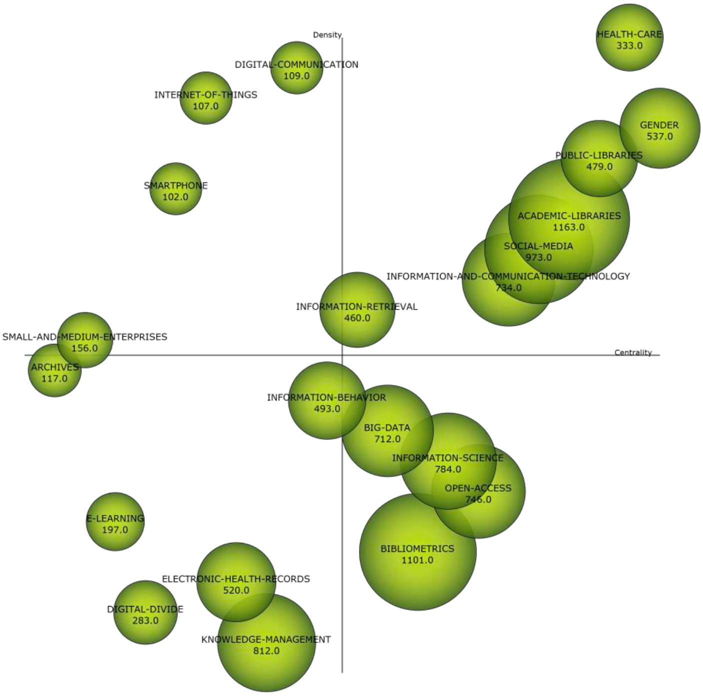


Figure 2. Local collaboration strategic diagram.

As we can see in Figure 2 and Table 2, the local collaboration is focused on some of the following themes. The theme *academic libraries*, which is related to information literacy, collaboration, and higher education. *Bibliometrics*, which is about citation analysis, altmetrics, and social network analysis. Additionally, we can see *social media*, which is delved into topics such as social networks, social networking sites, and political communication. Also, new technologies such as the *big data* theme have a place in this collaboration type, focusing on concepts such as e-government, transparency, open data, and developing countries. Another important theme is *information and communication technology* that is related to content analysis, digital media, and journalism.

The themes with the strongest impact are *internet of things* that is delved into digitalization, academic librarians, and blockchain, which is a new technology too; and *digital divide*, which is related to cloud computing, user experience and customer relationship management, obtains a high impact in this collaboration type. These last two themes have a MNCS of 2.32 and 1.41 respectively, as shown in Table 2. In contrast, the themes that achieve less impact are *academic libraries* and *information science* with a MNCS of 0.63 and 0.68, respectively.

Table 2. Local collaboration cluster measures.

Cluster names	*#Docs*	*Citations average*	*h-Index*	*MNCS*
ACADEMIC-LIBRARIES	1,163	5.44	29	0.63
BIBLIOMETRICS	1,101	10.03	41	1.14
SOCIAL-MEDIA	973	10.23	42	1.25
KNOWLEDGE-MANAGEMENT	812	9.27	35	1.14
INFORMATION-SCIENCE	784	5.25	25	0.68
OPEN-ACCESS	746	6.61	27	0.78
INFORMATION-AND-COMMUNICATION-TECHNOLOGY	734	6.63	29	0.8
BIG-DATA	712	9.69	37	1.12
GENDER	537	6.67	23	0.86
ELECTRONIC-HEALTH-RECORDS	520	9.87	29	1.29
INFORMATION-BEHAVIOR	493	5.36	20	0.72
PUBLIC-LIBRARIES	479	6.37	24	0.75
INFORMATION-RETRIEVAL	460	7.96	20	0.89
HEALTH-CARE	333	7.15	22	0.86
DIGITAL-DIVIDE	283	12.21	27	1.41
E-LEARNING	197	8.85	19	1.13
SMALL-AND-MEDIUM-ENTERPRISES	156	8.26	17	1.06
ARCHIVES	117	4.91	13	0.67
DIGITAL-COMMUNICATION	109	10.25	13	1.22
INTERNET-OF-THINGS	107	13.36	18	2.32
SMARTPHONE	102	7.52	14	0.91

Furthermore, the national collaboration, as shown in Figure 3 and Table 3, puts its attention on the following themes: *bibliometrics*, which is focused on the same topics as the local collaboration type; *electronic health records*, which add new technologies to the health context such as machine learning and natural language processing; *academic libraries*, which is related to the same concepts as the local collaboration; and *social media*, which is focused on the same topics as in the local collaboration too. The stronger themes in terms of impact are *electronic* *health records* and *patient portals* with a MNCS of 1.56 and 1.63 respectively, as shown in Table 3. *Patient portals* theme is delved into mobile health and electronic health. Otherwise, the themes with lower impact are *academic libraries* and *ontology*. The *ontology* theme is related to information retrieval, data sharing and interoperability topics.

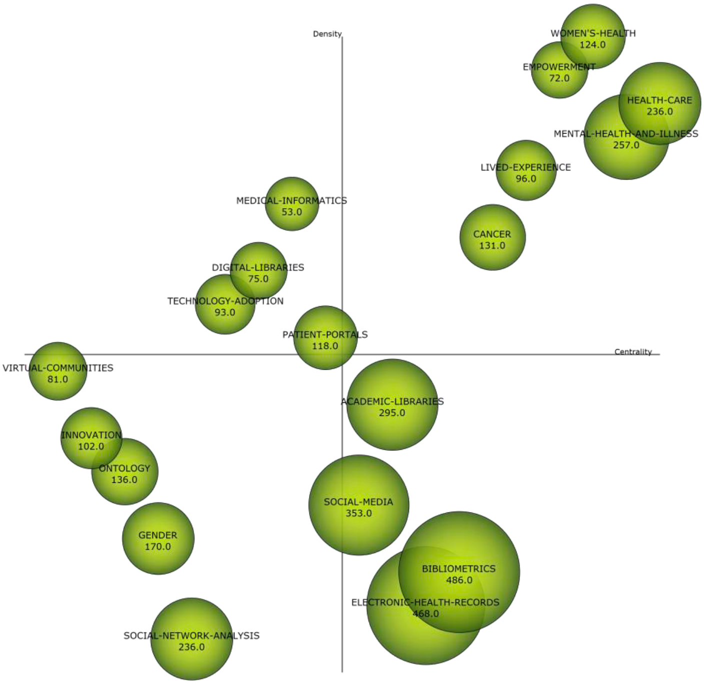


Figure 3. National collaboration strategic diagram

Moving to the international collaboration side, as shown in Figure 4 and Table 4, the themes have some similarities but also many differences with the other collaboration types: *bibliometrics*, but this time it is very related to data mining, content analysis and text mining; *developing countries*, focusing on topics such as information and communication technology and small and medium enterprises; *social media*, which relates to the same concepts as the other collaboration types; *knowledge management*, which is delved into knowledge sharing, innovation and systematic literature review; and themes focused on new technologies like *sentiment analysis*, *big data*, *natural language processing* and *machine learning*.

The theme with the stronger impact is the *literature review,* which is about information system and enterprise architecture, and obtains a MNCS of 3.13, as shown in Table 4. Also, the new technologies themes obtain a high impact. In contrast, more traditional themes like *higher education* and *health care* obtain a lower rate of MNCS, 0.97 and 1.27, respectively.

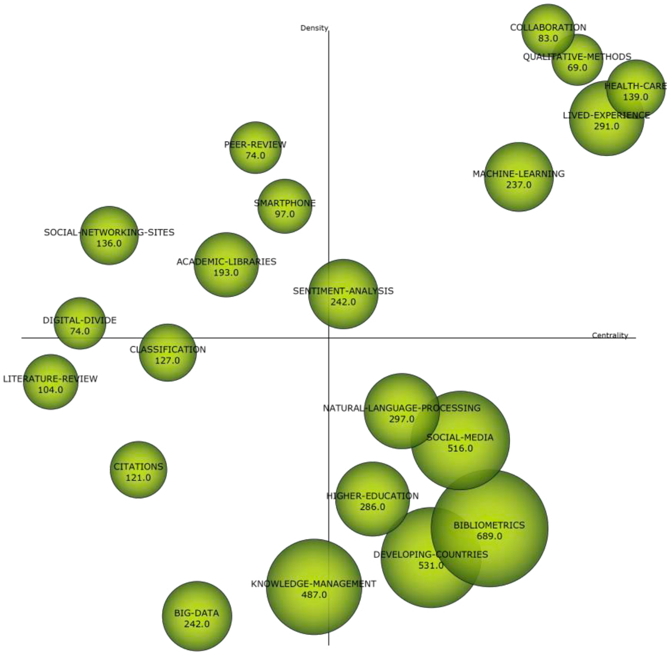


Figure 4. International collaboration strategic diagram

Table 3. National collaboration clusters measures

Cluster names	*#Docs*	*Citations average*	*h-Index*	*MNCS*
BIBLIOMETRICS	486	8.39	26	1.04
ELECTRONIC-HEALTH-RECORDS	468	11.35	33	1.56
SOCIAL-MEDIA	353	11.78	29	1.52
ACADEMIC-LIBRARIES	295	5.27	17	0.62
MENTAL-HEALTH-AND-ILLNESS	257	8.1	19	0.98
SOCIAL-NETWORK-ANALYSIS	236	10.15	25	1.29
HEALTH-CARE	236	9.52	23	1.19
GENDER	170	7.97	19	0.99
ONTOLOGY	136	6.64	17	0.78
CANCER	131	7.26	16	0.99
WOMEN'S-HEALTH	124	8.84	15	1.08
PATIENT-PORTALS	118	12.76	21	1.63
INNOVATION	102	11.69	17	1.39
LIVED-EXPERIENCE	96	8.98	16	1.06
TECHNOLOGY-ADOPTION	93	12.22	20	1.49
VIRTUAL-COMMUNITIES	81	11.84	17	1.19
DIGITAL-LIBRARIES	75	7.67	12	0.97
EMPOWERMENT	72	8.58	14	1.21
MEDICAL-INFORMATICS	53	11.43	13	1.5

Table 4. International collaboration clusters measures

Cluster names	*#Docs*	*Citations average*	*h-Index*	*MNCS*
BIBLIOMETRICS	689	11.36	39	1.38
DEVELOPING-COUNTRIES	531	12.36	37	1.83
SOCIAL-MEDIA	516	16.6	44	2.04
KNOWLEDGE-MANAGEMENT	487	13.27	37	1.85
NATURAL-LANGUAGE-PROCESSING	297	15.74	35	2
LIVED-EXPERIENCE	291	11.72	25	1.49
HIGHER-EDUCATION	286	7.73	23	0.97
SENTIMENT-ANALYSIS	242	20.54	37	2.51
BIG-DATA	242	16.26	33	2.34
MACHINE-LEARNING	237	12.26	21	1.62
ACADEMIC-LIBRARIES	193	8.93	23	1.23
HEALTH-CARE	139	9.74	18	1.27
SOCIAL-NETWORKING-SITES	136	17.35	27	2.13
CLASSIFICATION	127	9.83	20	1.29
CITATIONS	121	9.29	18	1.23
LITERATURE-REVIEW	104	22.61	25	3.13
SMARTPHONE	97	11.73	19	1.56
COLLABORATION	83	18.53	17	1.98
DIGITAL-DIVIDE	74	9.03	14	1.1
PEER-REVIEW	74	13.45	16	1.65
QUALITATIVE-METHODS	69	6.14	12	1.1

Thematical similarities and differences

Aiming to study the differences and similarities of the themes treated in each collaboration type, the results of the most cited keywords overlapping in terms of the Jaccard index are shown in Figure 5. The computation was done for the *n* most cited keywords, for exponentially increasing values of *n* until 10^7, which includes all the keywords in the datasets.

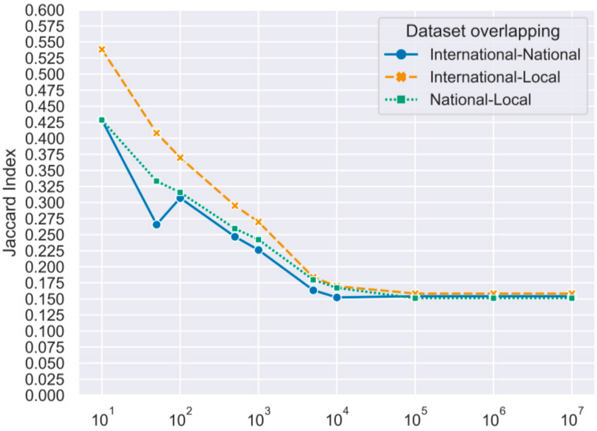


Figure 5. Jaccard Index between datasets keywords for n in 10, 50, 100, 500, 1000, 5000, 10000, 100000, 100000

As Figure 5 reveals, there is a moderate sharing of the most cited keywords between the international and local collaboration that is in line with the strategic diagrams obtained from the science mapping analysis. Moreover, the national and international collaboration is less shared than the international-local. Finally, the national and local collaboration share fewer keywords than the other ones, but it has a similar slope that is not as pronounced as the others. Therefore, the international and local collaboration are the most similar in terms of most cited keywords than the others.

Discussion

Once the proposed analysis has been done, an overview of the impact and the thematical differences and similarities between the collaboration types of the LIS category is given. First, according to previous studies (Gazni et al., 2012; Persson, 2010; Sin, 2011), international collaboration yields a higher impact in terms of citations. It is also demonstrated in our study by the results of the geometric mean (4.14 < 5.13 < 6.3) for local, national, and international collaboration, respectively. This fact is also supported because 25% of most cited papers of each collaboration type are more cited when researchers collaborate with international colleagues.

Regarding the research conducted, there are some thematical similarities as well as many differences between collaboration types. The similarities that can be noticed are: *bibliometrics*, which is present on the three collaboration types and obtains a similar impact in the three. It is delved into citation analysis, altmetrics and social network; moreover, *social media* theme is quite similar, and it is very related to social networks, but in the international collaboration it is a stronger theme in terms of MNCS; *health care* is very similar too, focusing on topics such as education, culture, and patient safety. In fact, there are several health-related themes in the three collaboration types: *empowerment*, which is about physical activity and mental illness; *gender*, which is about how gender relates to health, race, and information technology. All of them get around the same impact at the three collaboration types. Furthermore, *academic libraries*, which is related to collaboration and information literacy, is similar too. Nevertheless, in the international sphere, computer science (data mining, text mining) plays an important role. In fact, the MNCS of *academic libraries* is shot up in the international collaboration type, and doubles the impact acquired in the local and national collaboration types.

Additionally, there are many differences. While in the local and national collaboration, artificial intelligence topics are more spread through the themes, in the international collaboration type they form four themes: *natural language processing*, *big data*, *sentiment analysis* and *machine learning*, with a really high MNCS, almost all of them doubling the average citation rate of the area. Moreover, *literature review*, which is in the international collaboration type, is delved into enterprise architecture and information systems and it has the highest impact (MNCS = 3.13). Also, there are other themes with a significant impact: *social networking sites* and *collaboration*. However, these two themes are very spread in other themes at the local and national collaboration types.

The differences concluded in the last paragraphs are confirmed by the study of the overlapping keywords. The international and local collaboration datasets share less than 2/3 of 10 highest cited keywords. If this number is exponentially increased by the power of 10, the sharing quickly decreases to a value around 1/3 and then converges to 1/6 approximately for all the keywords. The rest of the collaboration datasets comparisons suffer the same situation, but they are less shared and converge to near the same value. This fact supports our performance analysis on the strategic diagrams which pointed out a difference between the themes of the different collaboration networks.

Although the findings reported in this contribution highlights the differences between the collaboration types, some limitations should be highlighted: only one database (Web of Science) was considered; only the period 2015-2019 was included. To the extent of our study, future research must include other databases and a wide period to reach a larger number of documents.

Conclusions

In this contribution, we present a research about the thematical differences and similarities in the Library and Information Science category between the 2015 and 2019 years (both included). Our study reveals that there are impact differences when a researcher collaborates with international colleagues rather than limiting the collaboration to the same institution or collaborating with colleagues from the same country. In fact, the more collaboration, the more impact is achieved in general. This fact has been pointed out and demonstrated in other areas and temporal periods (Chinchilla-Rodríguez et al., 2019; Gabrielle Breugelmans et al., 2018; Gazni et al., 2012; Polyakov et al., 2017). Furthermore, when this difference is detected, there is a difference between the themes in the collaboration types. The difference is mainly related to artificial intelligence being applied broadly in the field as well as *literature review*, *collaboration*, and *social networking sites*, acquiring a very high attention. However, there are also some similarities like *social media*, *health care* and *bibliometrics* themes. Nonetheless, while this difference in impact has been found in our study and we have analyzed the thematical differences, it also can be due to external factors, such as publication place, funding (Asubiaro, 2019), research leadership (Chinchilla-Rodríguez et al., 2019) or publishing in open access (Gabrielle Breugelmans et al., 2018). Therefore, future research on this topic should consider these other variables in order to demonstrate what we are stating in this study.

Acknowledgments

This work has been supported by the Spanish State Research Agency through the project PID2019-105381GA-I00 / AEI / 10.13039/501100011033 (iScience).

References

Adams, J. (2013). The fourth age of research. *Nature*.

Asubiaro, T. (2019). How collaboration type, publication place, funding and author's role affect citations received by publications from Africa: A bibliometric study of LIS research from 1996 to 2015. *Scientometrics*, *120*(3), 1261–1287. Springer Netherlands.

Batagelj, V., & Cerinšek, M. (2013). On bibliographic networks. *Scientometrics*.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*.

Chinchilla-Rodríguez, Z., Sugimoto, C. R., & Larivière, V. (2019). Follow the leader: On the relationship between leadership and scholarly impact in international collaborations. *PLOS ONE*, *14*(6), 1–18. Public Library of Science. Retrieved from https://doi.org/10.1371/journal.pone.0218309

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011a). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011b). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *Journal of Informetrics*, *5*(1), 146–166.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, *63*(8), 1609–1630. Retrieved from http://doi.wiley.com/10.1002/asi.22688

Van Eck, N. J., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology*.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., et al. (2018). Science of science. *Science*.

Gabrielle Breugelmans, J., Roberge, G., Tippett, C., Durning, M., Struck, D. B., & Makanga, M. M. (2018). Scientific impact increases when researchers publish in open access and international collaboration: A bibliometric analysis on poverty-related disease papers. *PLoS ONE*.

Galvez, C. (2018). Co-word analysis applied to highly cited papers in Library and Information Science (2007-2017). *Transinformacao*, *30*(3), 277–286.

Gazni, A., Sugimoto, C. R., & Didegah, F. (2012). Mapping world scientific collaboration: Authors, institutions, and countries. *Journal of the American Society for Information Science and Technology*.

Hu, C.-P., Hu, J.-M., Deng, S.-L., & Liu, Y. (2013). A co-word analysis of library and information science in China. *Scientometrics*, *97*(2), 369–382.

Khor, K. A., & Yu, L. G. (2016). Influence of international co-authorship on the research citation impact of young universities. *Scientometrics*.

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*.

Leydesdorff, L., Bornmann, L., & Wagner, C. S. (2019). The Relative Influences of Government Funding and International Collaboration on Citation Impact. *Journal of the Association for Information Science and Technology*.

Milojević, S. (2015). Quantifying the cognitive extent of science. *Journal of Informetrics*, *9*(4), 962–973.

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *El Profesional de la Información*, *29*(1). Ediciones Profesionales de la Informacion SL.

Olmeda-Gómez, C., Ovalle-Perandones, M.-A., & Perianes-Rodríguez, A. (2017). Co-word analysis and thematic landscapes in Spanish information science literature, 1985–2014. *Scientometrics*, *113*(1), 195–217.

Persson, O. (2010). Are highly cited papers more international? *Scientometrics*.

Polyakov, M., Polyakov, S., & Iftekhar, M. S. (2017). Does academic collaboration equally benefit impact of research across topics? The case of agricultural, resource, environmental and ecological economics. *Scientometrics*.

Rousseau, R., & Ding, J. (2016). Does international collaboration yield a higher citation potential for US scientists publishing in highly visible interdisciplinary Journals? *Journal of the Association for Information Science and Technology*.

Sin, S. C. J. (2011). International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980-2008. *Journal of the American Society for Information Science and Technology*.

Sooryamoorthy, R. (2017). Do types of collaboration change citation? A scientometric analysis of social science publications in South Africa. *Scientometrics*.

Sugimoto, C. R., Robinson-Garcia, N., Murray, D. S., Yegros-Yegros, A., Costas, R., & Larivière, V. (2017). Scientists have most impact when they're free to move. *Nature*.

Suresh, S. (2012). Global challenges need global solutions. *Nature*, *490*(7420), 337–338.

Thelwall, M. (2016). The precision of the arithmetic mean, geometric mean and percentiles for citation data: An experimental simulation modelling approach. *Journal of Informetrics*, *10*(1), 110–123. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S1751157715301437

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1). Nature Publishing Group.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47.

Assessing the potential effect of a unified system of journal classification based on citation impact indicators: the new Qualis

Eloisa Viggiani[1], Alexandre Prestes Uchoa[2] Jano Moreira Souza[2]

[1]*celeloisa@gmail.com*
Universidade Federal do Rio Grande do Sul – UFRGS, Rio de Janeiro (Brazil)

[2]*{auchoa, jano}@cos.ufrj.br*
Universidade Federal do Rio de Janeiro – UFRJ, Rio de Janeiro (Brazil)

Abstract

In 2018, CAPES ' Evaluation Directorate proposed a new methodology for Qualis, the journal classification system used to score the scholarly output of Masters and PhD programmes in Brazil, which used a system with varied composite criteria specific to each Evaluation Area. A unified criterion without distinction by areas was proposed, based on WoS, Scopus and GS citation impact indicators. We explored the potential effect on the publication scores of programmes and Evaluation Areas if a such a new system was adopted and, through correlational analysis using various characteristics of journals and programmes, we found signs that the reasons for the negative effects may have different origins and, therefore, can be mitigated through different actions. We found that the Sports Education, Nursing, and Agricultural Sciences areas would see the strongest decrease in their publication scores and that at least 96% of their programmes would be negatively affected. Journals published by Brazilian universities and societies, which may have important functions for the programmes, such as training of early career researchers, were the most downgraded. Our study brings insights to the improvement of the proposed classification system, and to the publishing strategy of programmes.

Introduction

The problem of assessing research output from different subject fields has long been studied in the context of Performance-based Research Funding Systems (PRFS) (Hicks, 2012) and the reliability and validity of citations as performance indicators has been debated for the past decades (Aksnes, Langfeldt and Woulters, 2019). Citation indicators and other bibliometric methods are believed to provide benefits in terms of robustness, validity, functionality, cost, and time of execution, when applied to the natural and formal sciences (Abramo and D´Angelo, 2011), though it is recognised that journals that are not covered by the main international bibliographic databases may have important functions for certain research fields and local research communities, such as knowledge gap-filling, knowledge bridging, introduction of new concepts and methods, and training of early career researchers, (Chavarro and Ràfols, 2019).

Many PRFS have addressed the issue by using a formula of weighted publications according to the rank or citation impact of the publication channel, such as the Norwegian Model, which attempts to comprehensively cover all the peer-reviewed scholarly journals in all areas of research in one single indicator, and where the bibliographic data are delivered by the institutions themselves through a CRIS system (Sivertsen, 2018). Nevertheless, PRFS are not static, and many have undergone major redesign (Hicks, 2012).

Other main research strands around the use of indicators in evaluations study the potential effects of evaluation exercises on publication practices, such as strategic behaviour by research communities, goal displacement, "gaming" of indicators, and the potential biases against interdisciplinary research (de Rijcke et al., 2016).

Background and purpose

Changes to the National Evaluation of Graduate Programmes in Brazil

The national evaluation of the graduate programmes in Brazil is performed by the Federal Agency for Support and Evaluation of Graduate Education (CAPES) every 4 years, and is a metrics-informed peer review process, through Evaluation Area committees, currently divided into 49 Areas (Oliveira and Amaral, 2017).

In 2018, CAPES ' Evaluation Directorate initiated a series of actions aiming to improve the instruments used in the evaluation. The main motivation was to increase the focus on the quality of the training of doctors and masters and on the excellence of Brazilian graduate studies. A new framework for multidimensional assessment was proposed, considering: i) teaching and learning, ii) internationalization, iii) knowledge production, iv) innovation and knowledge transfer, and v) economic and social impact, and relevance. Working groups started in 2019 to elaborate the definition of concepts, variables and indicators representing each of the evaluation dimensions (CAPES, 2020).

Of special interest to our study is the proposed methodology for a unified Qualis journal classification for all Evaluation Areas, hereafter identified as “Qualis 2019”. We will briefly describe the former Qualis system, the proposed changes, and their justification. Nevertheless, it is important to highlight that many other indicators are taken into consideration in the overall evaluation framework, related to other functions that Master and PhD programmes need to fulfil.

The former Qualis System

The Qualis journal classification system was created in 1988 to allow the analysis and evaluation of the scholarly output reported by degree programmes in all disciplinary areas. Until the last national evaluation exercise performed in 2017, covering the period from 2013 to 2016, the criteria for journal classification was determined by each Evaluation Area Committee. The Committees were also responsible for continuously updating their Qualis journal list, and minor revisions to the criteria were not uncommon.

A recent study by Oliveira and Amaral (2017) found that 81% of the Evaluation Areas used the Impact Factor and 56% made use of SJR/Scimago. Many Areas classified journals based on their presence in regional or field specific databases, such as SciELO, RedALyC, LatIndex, PubMed, PSYCINFO, and BIOSIS. The Evaluation Areas also considered titles which were evaluated for their local or regional impact, the importance of the content developed and the objectives of the research. For example, within the area of Agricultural Sciences, the development of literature, teaching of local history among others, in which the less dissemination does not mean lack of quality or prestige (Oliveira and Amaral, 2017). This system, from now on identified as “Qualis 2017”, resulted in 49 different lists of journals and corresponding strata, published by CAPES as a web page with multiple filters (“Plataforma Sucupira”, [n.d.]).

The main objections to Qualis 2017 were around the concentration of publications in low visibility/low impact journals, the difficulty of comparisons and benchmarking, and management issues. Pires et al. (2020) argued that as an effect of the Qualis system used until 2017, Brazilian degree programmes concentrated their publications in a restricted number of journals, which were classified in the top levels of Qualis, regardless of their performance in terms of international visibility and citation impact. Perlin, Imasato, and Borenstein (2018) found that when a potentially predatory journal enters Qualis, the number of publications from Brazilian programmes tends to increase, meaning that researchers are less likely to challenge its quality and visibility. The authors also found that even experienced researchers, with significant career time and many publications in non-indexed journals are publishing in questionable journals. According to Barata (2016), the system produced incommensurable results, hampering the integration between the Evaluation Areas, and resulting in lack of communication between key stakeholders. The author justified the need for a unified Qualis classification system as the only way to give greater visibility to Brazil´s scientific production and encourage the publication of quality research in all areas of knowledge.

The new Qualis System

In this new system proposed by CAPES in 2019, each journal would receive only one classification, based on field normalized citation impact percentiles, calculated using indicators from WoS (Impact Factor) or Scopus (CiteScore). The first step would be to assign each journal to a subject field (or "Mother Area") corresponding to the Evaluation Area whose programmes reported the highest number of publications during the evaluation period, and then consult the corresponding indicators. When both indicators from WoS and Scopus are available for a journal, the highest percentile value among them is to be considered. When neither is available, the value of Google Scholar´s h5 index is used in a model of regression making a relationship between h5 and CiteScore and estimating a corresponding percentile value. (CAPES, 2019a). Based on this method, a journal list was created, with all the titles previously classified in Qualis 2017, and their corresponding classification when using Qualis 2019. This list was presented to the Evaluation Areas Committees during the mid-term review in 2019 for their input and feedback. Currently, there is a working group reviewing this list and the final version is expected to be published in 2021.

The purpose of this study is to explore the Evaluation Areas (EAs) which would potentially be affected by the replacement of a journal classification system with distinctions between disciplinary fields (Qualis 2017) by a unified classification system based on citation impact indicators (Qualis 2019). Our goal is to identify characteristics of the main downgraded journals and of the programmes which would see a decrease in their publication score. Though acknowledging that a pure bibliometric approach is not sufficient to assess the suitability of the proposed unified system, our study aims to contribute to its refinement and improvement, and to support programmes in reviewing and/or adapting their publication strategies.

Methodology

Data Sources

The source of metadata regarding the programmes and their scholarly output were produced by CAPES' Sucupira system and are openly available in CAPES' open data repository (http://dadosabertos.capes.gov.br). Each register contains the Evaluation Area, programme name and identification number, programme degree level (Academic Master, Professional Master, Academic Master and PhD, or Academic PhD), programme start year, grade received in the 2017 evaluation exercise, institution, number of publications, publication year, journal name, and journal ISSN. The retrieved dataset contains 714,459 registers for the years 2017 and 2018, which reported a total of 1,245,755 publications in 16,272 unique journals, authored by faculty and students from 4,442 programmes, offered by 246 Brazilian institutions. The criteria for journal classification adopted by each EA were also obtained from Sucupira System. The Qualis 2017 journal classifications were published by CAPES as a web page with multiple filters ("Plataforma Sucupira", [n.d.]). As a result, only isolated parts of the classification could be consulted at a time. Additionally, there was no print or download feature. For practical reasons, we decided to use the files released by (GOUVEIA, 2017). This dataset consists of independent CSV files, each one for an Evaluation Area with all its corresponding journals and respective strata. The Qualis 2019 journal classification, although not officially released yet, could be easily found in the websites of many universities as a single text file. After comparing the content of the files of three different websites and making sure they were identical, their content was converted into a table.

Qualis 2017 and Qualis 2019 strata and scoring systems

In the Qualis 2019 system there are 8 strata, corresponding to one citation impact octile each, and the scoring is equally divided into the strata. In the Qualis 2017 system, there were 7 strata, with a rule limiting the number of journals classified in the upper strata, as can be seen in Table 1, which contains the classification strata, scores, and journal distribution quartiles for Qualis 2017 and Qualis 2019.

Table 1. Qualis 2017 and Qualis 2019 classification strata, scores, and corresponding journal distribution quartiles.

Qualis 2017		*Journal Distribution Quartiles*	*Qualis 2019*	
Score	Strata		Strata	Score
1.000	A1	Q1	A1	1.000
0.850	A2		A2	0.875
0.700	B1	Q2	A3	0.750
			A4	0.625
0.550	B2	Q3	B1	0.500
0.400	B3		B2	0.375
0.250	B4	Q4	B3	0.250
0.100	B5		B4	0.125
0	C		C / NP	0

Identification of most affected Evaluation Areas and Programmes

A total publication score was calculated for each Evaluation Area and each programme, using both the Qualis 2019 and the Qualis 2017 scoring system, and their difference was used to determine which Evaluation Areas and which programmes would be more positively or negatively affected by the new classification system. We selected the three most negatively affected Evaluation Areas (EAs) for the following analysis.

Analysis of Qualis 2017 criteria and journal up- or downgrade

The 2017 Area Documents were consulted to verify the criteria for journal classification adopted by the EAs. The criteria were categorised as follows:

- **CI** = used citation impact indicators from WoS, Scopus and/or Google Scholar.
- **COV** = considered the coverage by databases, such as: SciELO, PubMed, LatIndex, and RedALyC. This category also included the coverage by WoS, Scopus or Google Scholar, when no citation impact indicators were used.
- **DF** = criteria applied only to journals with a selected disciplinary focus or gave those journals some advantage over multidisciplinary journals and journals considered from unrelated disciplines.
- **BR** = criteria applied only to Brazilian journals or gave those journals some advantage over the others.

For each EA, we calculated the percentage of journals and publications which continued to be classified in the same quartile and of those which were upgraded or downgraded, when using the Qualis 2019 classification. The average number of articles per title was calculated to identify if there was a higher concentration of articles on the titles that remained in the same quartile or on those up- or downgraded.

Characteristics of main downgraded journals

For all journals which published 10 or more articles from the selected Evaluation Areas in 2017 and 2018, namely the "main downgraded journals", we identified the journal´s publisher, country, and type of organisation (university, scientific society, or commercial publisher), and calculated the percentage of journals and publications published by:

- Brazilian universities, scientific societies, and commercial publishers, which potentially have the function of knowledge gap-filling, knowledge bridging and introducing new concepts and methods to local communities of research and practice.
- Brazilian universities, which could also have the function of training early career researchers.

Since one of the most affected Evaluation Areas reported the presence of potentially predatory journals in their Qualis 2017 list, we also calculated the percentage of their publications.

Identification of Programmes Characteristics

We investigated various characteristics in search of possible reasons for the publication practices that, in the final analysis, would lead to low scores in case of using exclusively a citation impact indicator. Having in mind that each Evaluation Area has its own characteristics inherent to its scientific field, our strategy focused on aspects related to the time of existence, experience, degree of multi or interdisciplinarity of teams, and whether these could be related to decreases in the score of publications. Regarding the programme´s experience, we verified the percentage of faculty graduated over 20 years ago and calculated the programme´s time of existence (years), with the following age quartiles:

- Q1 = programmes with up to 2 years of existence
- Q2 = between 2 and 7 years
- Q3 = between 7 and 13 years
- Q4 = more than 13 years

The programme´s multidisciplinarity was assessed by verifying the percentage of faculty graduated in a programme belonging to a different Evaluation Area. The balance was assessed by the grade received by the programme in the last evaluation exercise. It is worth mentioning that this was considered an especially useful measure of balance, because despite being an aggregation which includes the publications, the programme grade also considers other important indicators of education, research, and third mission, such as number of students per faculty, theses, dissertations, and patents. Grades range from 1 to 7, where 1 is the lowest and 7 is the highest grade. Other contextual characteristics were also analysed, such as number of faculty members, proportion of permanent vs. visiting faculty, and geographic region (in Brazil), but no significant relation to the decrease in publication scores was found.

Findings

Most affected Evaluation Areas

Out of the 49 Evaluation Areas, 15 would see a decrease in their average publication score if the new Qualis 2019 system was implemented, affecting more than 50% of their programmes. The most negatively affected Evaluation Area would be Sports Education, regarding both the decrease in the average publication score (-2318), as well as in the proportion of programmes which had their publications scores decreased (100%), followed by Nursing (-1594, 99%), and Agricultural Sciences (-1700, 96%). Figure 1 shows the average difference in the Evaluation Area publication score and the percentage of programmes which would have a decrease in their publication score, when using Qualis 2019 and Qualis 2017, for the 49 Evaluation Areas.

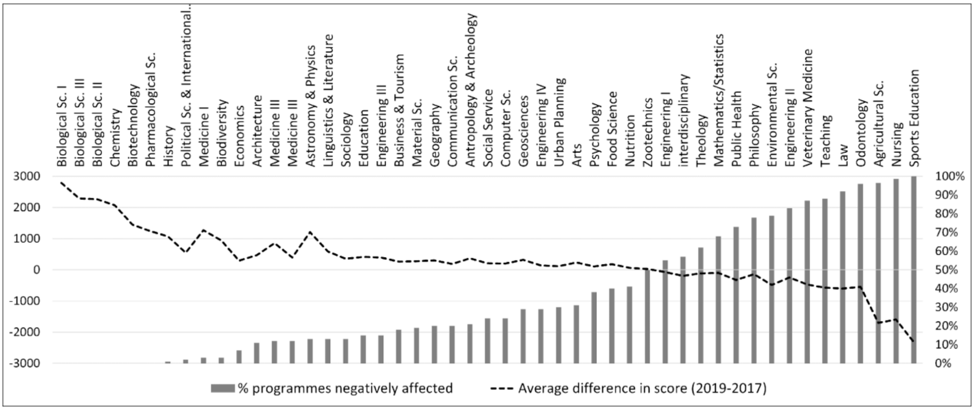


Figure 1: Average difference in the publication score of the Evaluation Areas when using Qualis 2019 and Qualis 2017, and the percentage of their programmes which would have a decrease in their publication score with the new system.

We selected the 3 most negatively affected Evaluation Areas for further analysis: Sports Education, Nursing, and Agricultural Sciences.

Evaluation Area Sports Education

The criteria for journal classification adopted by the Sports Education Area for the Qualis 2017 used impact citation indicators from WoS (SCIE and SSCI) for the 1st and 2nd journal quartiles, but only for journals that met the EA´s disciplinary focus, which included the subject categories “Sport Sciences”, “Rehabilitation”, “Audiology & Speech-Language Pathology”, “Hospitality, Leisure, and “Sport & Tourism” in WoS. The highest possible classification for journals from other disciplines was in the 2nd quartile. For the 3rd and 4th journal quartiles, the Sports Education area opted for a combination of coverage by SciELO, Pubmed, CINAHL, RedALyC and/or LatIndex. As expected, due to the different criteria, we found that a significant proportion of journals would be upgraded, as well as downgraded, if Qualis 2019 was adopted, as can be seen in Table 2.

Table 2. Qualis 2017 criteria for the Evaluation Area Sports Education and journal classification by Qualis 2017 and Qualis 2019.

Journal Quartile	*Qualis 2017 Criteria*	*Journals classified by Qualis 2017*	*Journals classified by Qualis 2019*		
			Upgraded	*In same quartile*	*downgraded*
Q1	CI+DF	281	-	178 (63.3%)	103 (36.7%)
Q2	CI+DF	182	53 (29.1%)	69 (37.9%)	60 (33.0%)
Q3	COV	215	110 (51.1%)	51 (23.7%)	54 (25.1%)
Q4	COV	190	69 (36.3%)	121 (63.7%)	-

CI = Citation impact indicators; DF = Disciplinary focus; COV = covered by database(s)

We found that the most significant reason for the decrease in the EA´s publication score is that the potentially downgraded journals are the ones which concentrated the highest number of publications from the Sports Education area. Even though the criteria adopted in 2017 did not include any advantage for Brazilian journals, we found that among the 10 journals with more publications from the Sports Education Area, 8 are Brazilian journals, and all would be downgraded with the adoption of Qualis 2019. We found 79 main downgraded journals responsible for 41.5% of the area´s output, of which 40 are Brazilian journals (2476 publications, 26.8%), predominantly published by universities (22 journals, 1410 publications, 15.2%) and societies (16 journals, 958 publications, 10.4%). Among the non-Brazilian journals, the majority of the titles belongs to commercial publishers (23 journals, 832 publications, 9%). Among these, we also found titles from the most prestigious international publishing houses, though they were the ones with the smallest decrease.

All 67 programmes belonging to Sports Education would see some degree of decrease in their publication score if Qualis 2019 was adopted and the strongest decreases would be among the 4 programmes which received grade 6 in the last evaluation, followed by the 2 programmes which received grade 7. These programmes represent only 11% of the total, so the decrease of the Evaluation Area´s total publication score would come mostly from the accumulated contribution of numerous programmes with lower grades, as seen in Figure 2 (left). Regarding the programme´s experience, the strongest decrease in publication scores was found in programmes with 7 to 13 years of existence, followed by programmes with more than 13 years, as seen in Figure 2 (right).

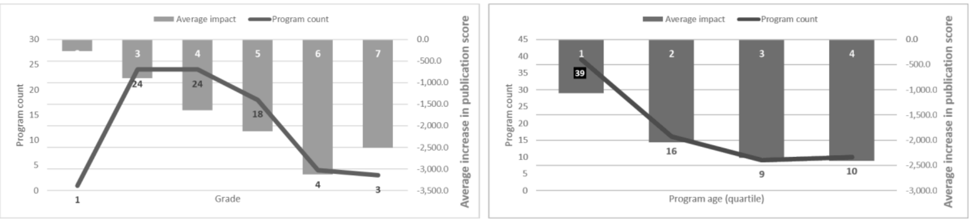


Figure 2: Relation between average decrease in programme´s publication score and grade received in the last evaluation exercise (left), and programme´s years of existence (right) for the Evaluation Area Sports Education. Lines represent number of programmes.

On the other hand, programmes in the Sports Education Area with more experienced faculty seem to be less affected than those with a higher proportion of faculty graduated more recently, as can be seen in Figure 3, which relates the average decrease in the programme´s publication score and the percentage of faculty graduated 20 years ago or more.

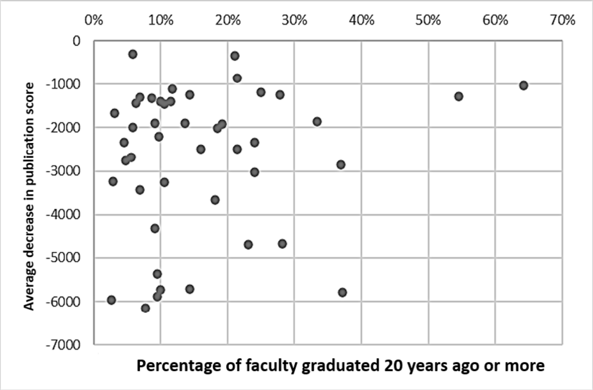


Figure 3: Average decrease in the publication score of programmes when using Qualis 2019 and percentage of faculty graduated 20 years ago or more for the Evaluation Area Sports Education.

Evaluation Area Nursing

The criteria for journal classification adopted by the Nursing Area for Qualis 2017 contemplated impact indicators from WoS, Scimago for the assignment of journals to the 1st quartile, but his criterion was only valid for journals in the subject category “Nursing” in WoS or Scopus. For the remaining quartiles, the Nursing Area considered the coverage by field specific databases, such as CUIDEN, a database published by Ciberindex, Medline, CINAHL, LILACS, BDENF, LatIndex, SciELO, and Rev@Enf.

Despite having adopted a similar criterion for the top journal quartile as in Qualis 2019, the Nursing Area would see 31.6% of journals in this quartile being downgraded with the adoption of the new classification system. An even higher proportion of journals would be downgraded from the 2nd quartile (58.7%) and from the 3rd quartile (44.2%), as can be seen in Table 3.

Table 3. Qualis 2017 criteria for the Evaluation Area Nursing and journal classification by Qualis 2017 and Qualis 2019.

Journal Quartile	*Qualis 2017 Criteria*	*Journals classified by Qualis 2017*	*Journals classified by Qualis 2019*		
			Upgraded	*In same quartile*	*downgraded*
Q1	CI+DF	155	-	106 (63.4%)	49 (31.6%)
Q2	COV	126	24 (19.0%)	28 (22.2%)	74 (58.7%)
Q3	COV	156	30 (19.2%)	57 (36.5%)	69 (44.2%)
Q4	COV	175	80 (45.7%)	95 (54.3%)	-

CI = Citation impact indicators; DF = Disciplinary focus; COV = covered by database(s)

Additionally, we found that the publications from the Nursing Area were highly concentrated in a small number of journals. No criterion to benefit Brazilian journals was adopted in Qualis 2017, but the main downgraded titles are predominantly published by Brazilian universities (9 journals, 1233 publications, 14.1%) and Brazilian societies (7 journals, 242 publications, 2.8%). 42.7% of the EA´s reported output was concentrated in 10 Brazilian journals, of which 5 would be downgraded by 1 or 2 quartiles.

Out of the 74 programmes in Nursing, 73 would have a decrease in their weighted publication scores and similarly to what was found in the Sports Education Area, the most significant decreases are again among the programmes which received higher grades in the last evaluation exercise, and among programmes with more than 13 years of age. Here again, the accumulated contribution of numerous programmes with lower grades has led to the overall decrease in the Evaluation Area´s total publication score, as seen in Figure 4 (left). However, differently from the Sports Education Area. However, we did not find a relation between the programme´s years of existence, which suggests that the decrease in the Nursing Area´s publication score is not affected by characteristic, as seen in Figure 4 (right).

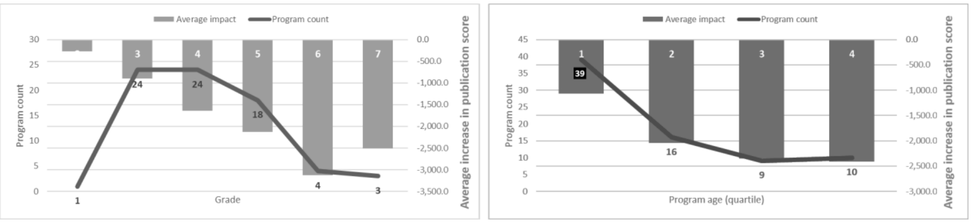


Figure 4: Relation between average decrease in programme´s publication score and grade received in the last evaluation exercise (left), and programme´s years of existence (right) for the Evaluation Area Nursing. Lines represent number of programmes.

A stronger relation was found regarding the multidisciplinary of the programmes´ faculty body. Programmes with a higher percentage of faculty members graduated in other disciplinary fields would be the ones less negatively affected by the adoption of Qualis 2019, possibly due to a selection of journals from other disciplinary fields, which may have been under classified by the Nursing Area in Qualis 2017. c

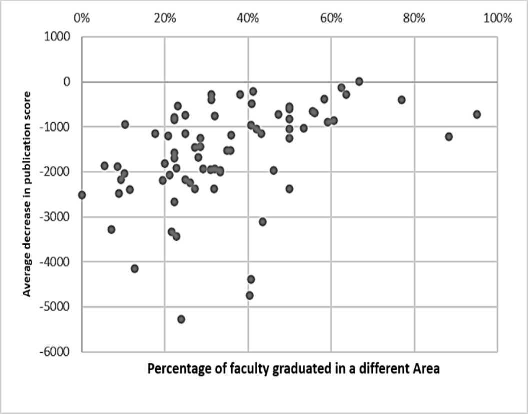


Figure 5: Average decrease in the publication score of programmes when using Qualis 2019 and percentage of faculty graduated in a different Area the Evaluation Area Nursing.

Evaluation Area Agricultural Sciences:

The Agricultural Sciences Area adopted citation impact indicators from WoS or Scopus for the classification of journals into the 1st and 2nd journal quartiles. For the 3rd and 4th quartiles, the EA considered the coverage by SciELO and the following field specific databases: CAB (Commonwealth Agricultural Bureau), BIOSIS (Biological Abstracts), and AGRIS (International System for Agricultural Science and Technology). We found that the majority of journals would remain in the same quartile with the adoption of Qualis 2019, as illustrated in Table 4.

Table 4. Qualis 2017 criteria for the Evaluation Area Agricultural Sciences and journal classification by Qualis 2017 and Qualis 2019.

Journal Quartile	*Qualis 2017 Criteria*	*Journals classified by Qualis 2017*	*Journals classified by Qualis 2019*		
			Upgraded	*In same quartile*	*downgraded*
Q1	CI	492	-	407 (82.7%)	85 (17.3%)
Q2	CI + DF	400	108 (27.0%)	183 (45.8%)	109 (27.3%)
Q3	CI + COV	264	79 (29.9%)	108 (40.9%)	77 (19.2%)
Q4	COV	474	160 (33.8%)	314 (66.2%)	-

CI = Citation impact indicators; DF = Disciplinary focus; COV = covered by database(s)

Nevertheless, in all journal distribution quartiles there was a significantly higher concentration of articles per titles among the downgraded journals, and this is the predominant factor for the decrease in the EA´s publication score. We found 89 main downgraded journals, with 7713 publications, representing 27.6% of the EA´s output for 2017 and 2018. Despite not having a criterion to benefit Brazilian journals in 2017, 33 main downgraded journals are published by Brazilian universities (3278 publications, 11.7%), and 17 by Brazilian societies (1924, 6.9%). 13 journals by non-Brazilian commercial publishers identified by the EA as potentially predatory published 1486 articles in 2017 and 2018, representing 5.3% of the area´s output, and may indeed be a source of concern for the EA.

Programmes with higher grades (5 to 7) in the last evaluation exercise would see the strongest decrease in their publication scores, as well as programmes with more than 13 years of existence, as can be seen in Figure 6.

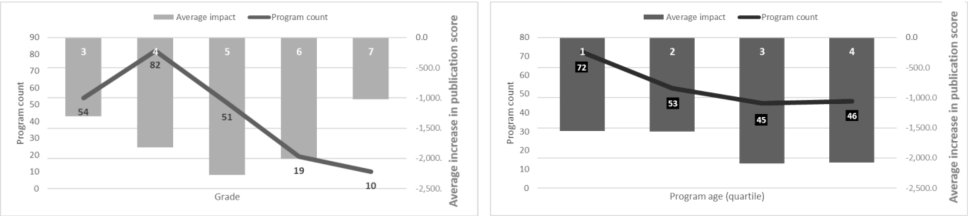


Figure 6: Relation between average decrease in programme´s publication score and grade received in the last evaluation exercise (left), and programme´s years of existence (right) for the Evaluation Area Agricultural Sciences. Lines represent number of programmes.

It is worth noting the high proportion of programmes which received high grades in the last evaluation exercise that would see a decrease in their publication scores with the adoption of Qualis 2019, and the degree in which this would happen. The same is valid for long standing programmes, especially those with more than 13 years of existence, and for programmes with well experienced faculty, which could be affected because of their selection of journals.

No meaningful relation was found between the potential decrease in the publication score of programmes in the Agricultural Sciences and either the percentage of faculty graduate in different Areas nor in the percentage of faculty graduated 20 years ago or more, as seen in Figure 7.

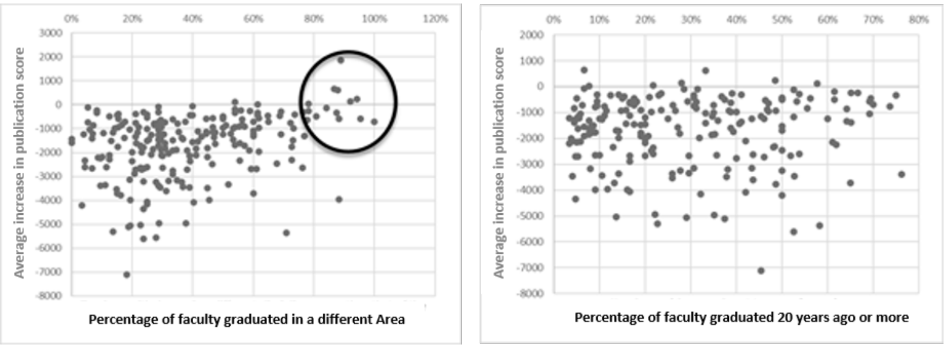


Figure 7: Average decrease in the publication score of programmes when using Qualis 2019 and percentage of faculty graduated in a different Area (left) and percentage of faculty graduated 20 years ago or more (right) for the Evaluation Area Agricultural Sciences. Each dot represents a programme. Programmes with no faculty graduated in a different Area were omitted.

Discussion and conclusions

The 3 Evaluation Areas which would experience the highest decrease in their publication scores if Qualis 2019 was adopted are Sports Education, Nursing, and Agricultural Sciences. These EAs used a combination of citation impact indicators from WoS and Scopus and the coverage by regional and/or field specific databases for journal classification in 2017, which resulted in significant changes in the journals´ classification strata and consequently, in the publication score of programmes with the adoption of a system based on citation impact from WoS, Scopus and GS. Nevertheless, in all 3 areas, part of the journals maintained their classification level in both systems, meaning that they have met both the bibliometric approach as well as their areas´ understanding of relevance and quality.

Despite not having adopted any criterion to benefit Brazilian journals, most of the potentially main downgraded journals are published by Brazilian universities and societies and may serve local communities by bringing new concepts and methods, knowledge bridging and knowledge gap-filling. Journals published by Brazilian universities, notably underfunded, may also have the function of training of early career researchers, which is part of the programmes´ educational mission. Further considerations are needed around what actions could be taken regarding these journals, especially because they concentrated the highest proportion of articles from the Evaluation Areas addressed in this study. If one understands that these journals serve specific needs of the programmes, the methodology for the new Qualis system could be adjusted for that purpose, although this means departing from a more universal and general system. Alternatively, strengthening these journals´ by means of directed resources and support, so they have a better chance to improve their editorial quality standards, could fulfil that goal. Although it is not guaranteed that by improving their quality they would consequently be indexed by the main bibliographic databases.

At the same time, this study showed signs that there may be programs that, even with an eventual adjustment in the categorization of certain journals, will still need to review their publication strategies. Further investigations are needed to understand why the journals chosen might be of relevance to the programmes and can precisely characterize each case to allow targeted, more effective, and reasonable actions like training of researchers and libraries on journals search, selection, and publishing options, targeted funding of submissions and subscriptions, even on article writing and translation.

Future studies can also focus on the upgraded journals, and the underlying reasons for many of them having been under classified in Qualis 2017, such as not meeting the stated disciplinary focus of the EA. It is important to understand the potential effects of any change on publication and research practices of each EA to ensure that its disciplinary (or multi- inter-transdisciplinary) goals are neither unintentionally distorted nor hindered by a new journal classification system only for the sake of simplicity.

In summary, multiple factors play a role in the publication pattern of the graduate programmes in Brazil and in the characteristics of local journals. Addressing them is important to ensure the robustness, validity, and functionality of the new Qualis system, and to maintain an alignment with the multidimensional approach proposed by CAPES, that considers teaching and learning, internationalization, knowledge production, innovation and knowledge transfer, economic and social impact, and relevance. Another important consideration is that the new Qualis which is expected to be finalised by mid-2021 will be used to evaluate the research output from 2017 to 2020 and so, whatever changes are implemented, the results should be interpreted with caution, since both researchers and programmes coordinators were working under the conditions of the previous criteria and will need time to adapt to the changes.

References

Abramo, G., D`Angelo, C.A. (2011). Evaluating research: from informed peer review to bibliometrics. *Scientometrics*, 87:499–514 DOI 10.1007/s11192-011-0352-7.

Aksnes, D., Langfeldt, L., Woulters, P. (2019). Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. *SAGE Open*. January 2019. doi:10.1177/2158244019829575

Barata, R.C. (2016). Ten things you should know about the Qualis. *Revista Brasileira de Pós-Graduação*, V.13-30, 13-40.

CAPES. (2016a). Considerações sobre o Qualis Periódicos – Enfermagem. Retrieved November 26, 2020 from: https://www.gov.br/capes/pt-br/centrais-de-conteudo/Enfermagem_Qualis_Peridicos__2016_revisado.pdf.

CAPES. (2019a). Aprimoramento do processo de avaliação da pós-graduação. Retrieved November 26, 2020 from: http://antigo.capes.gov.br/images/novo_portal/documentos/DAV/avaliacao/18072019_Esclarecimentos_Qualis2.pdf

CAPES. (2019c). Grupo de Trabalho Internacionalização. Relatório e Recomendações. Retrieved January 5, 2021 from https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/avaliacao/relatorios-tecnicos-e-grupos-de-trabalho.

CAPES. (2019d). Relatório do Seminário de Meio Termo – Área Enfermagem. Retrieved November 6, 2020 from https://www.gov.br/capes/pt-br/centrais-de-conteudo/documentos/avaliacao/RELATORIO_SEMINARIO_MEIO_TERMO_MP.pdf.

CAPES. (2020). Orientações sobre o processo avaliativo CAPES ciclo 2017-2020. Retrieved January 5, 2021 from: https://www.gov.br/capes/pt-br/centrais-de-conteudo/documentos/avaliacao/copy_of_ORIENTAES_PROCESSO_AVALIATIVO_INFORMATIVO_1.pdf.

Chavarro, D., Tang, P., Rafols, I. (2017) Why researchers publish in non-mainstream journals: Training, knowledge bridging, and gap filling. *Research Policy*. 46(9): 1666-1680. doi:10.1016/j.respol.2017.08.002.

de Rijcke, S. et al. (2016). Evaluation practices and effects of indicator use – a literature review. *Research Evaluation*, 25(2): 161-169.

Gouveia, F.C. (2017). Arquivos Qualis 2013-2016 de todas as áreas para ScriptLattesfigshare. Retrieved January 7, 2021 from https://figshare.com/articles/dataset/Arquivos_Qualis_de_todas_as_reas_para_ScriptLattes_2013-2016/5177185/2.

Perlin, M.S., Imasato, T., Borenstein, D. (2018). Is predatory publishing a real threat? Evidence from a large database study. *Scientometrics,* 116: 225-273.

Pires, A. S. et al. (2020). Implicações do sistema de classificação de periódicos Qualis em práticas de publicação no Brasil entre 2007 e 2016. *Arquivos Analíticos de Políticas Educativas*, 28(25).

Oliveira, T.M., Amaral, L. (2017). Public Policies in Science and Technology in Brazil: challenges and proposals for the use of indicators in evaluation. In ECA/USP (Ed.), *Bibliometrics and scientometrics in Brazil: scientific research assessment infrastructure in the era of Big Data* / Mugnaini, R. Fujino, A., Kobashi, N.Y. São Paulo, Available at: http://www.livrosabertos.sibi.usp.br/portaldelivrosUSP/catalog/download/129/108/547-1?inline=1

Rafols, I., Molas-Gallart, J., Chavarro, D. A., Robinson-Garcia, N. (2016) On the Dominance of Quantitative Evaluation in 'Peripheral' Countries: Auditing Research with Technologies of Distance (May 28, 2016). Available at: http://dx.doi.org/10.2139/ssrn.2818335

Sivertsen, G. (2018). The Norwegian Model in Norway. *Journal of Data and Information Science*. Vol. 3 No. 4, pp 2–18 DOI: 10.2478/jdis-2018-0017.

Measuring Funding Programs Achievements in Fostering Cross-Disciplinarity Research: The Potential Role of Integration and Collaboration Indicators in Capturing Distinct Support Mechanisms

Etienne Vignola-Gagné[1], Henrique Pinheiro[2], Maxime Rivest[3] and David Campbell[4]

[1] *e.vignola-gagne@elsevier.com*
Science-Metrix/Elsevier, 1335 Mont-Royal Ave E, Montreal, Quebec H2J 1Y6 (Canada)

[2] *h.pinheiro@elsevier.com*
Science-Metrix/Elsevier, 1335 Mont-Royal Ave E, Montreal, Quebec H2J 1Y6 (Canada)

[3] *m.rivest@elsevier.com*
Science-Metrix/Elsevier, 1335 Mont-Royal Ave E, Montreal, Quebec H2J 1Y6 (Canada)

[4] *d.campbell@elsevier.com*
Science-Metrix/Elsevier, 1335 Mont-Royal Ave E, Montreal, Quebec H2J 1Y6 (Canada)

Abstract

Twenty funding programs from many countries were selected for they cover, or not, a range of support mechanisms fueling, to varying degrees, cross-disciplinarity (XDR). Their publication outputs were then used to test the usefulness of quantitative indicators to capture their anticipated degree of XDR. This assumes that the programs have been sufficiently successful to produce XDR outputs to a degree that matches the qualitative assessment. Publication sets were delineated through grant databases or funding acknowledgements. The quantitative assessment relied on the use of two indicators of disciplinary diversity, 1) the Rao-Stirling index applied to references to measure intellectual integration (interdisciplinarity); and 2), a novel variation on the Rao-Stirling index applied to authorships (multidisciplinary collaboration). The quantitative results highlighted the current emphasis placed on multidisciplinary collaboration in the funding landscape for XDR, as programs producing XDR papers were scoring high either on multidisciplinary collaboration alone or on both dimensions, but never on interdisciplinarity alone. The XDR indicators were rarely above expectations for non-XDR programs, whereas they were mostly above, to varying degrees, for XDR programs. The results also showed that the XDR programs were either more or less successful than anticipated, or highlighted the challenge of making precise qualitative characterizations.

Introduction

Recent work has highlighted problems in the application of currently available bibliometric indicators of interdisciplinarity and cross-disciplinarity (here, we will use the term "cross-disciplinarity", abbreviated to XDR to encompass all modalities of interdisciplinarity, multidisciplinarity or related practices that are currently measurable with bibliometric indicators). Various teams have reported (Hackett et al., 2021; Leydesdorff & Bornmann, 2020; Locatelli et al., 2021) that findings from bibliometric indicators of XDR did not converge with observations from qualitative or case study work (Hackett et al., 2021). Schneider and Wang, reviewing multiple indicators of XDR, concluded that none of the currently available bibliometric approaches are fully satisfactory for capturing practices in XDR (Q. Wang & Schneider, 2019). The work by Locatelli et al (using mixed methods with bibliometrics being just one evidence stream considered) make it clear that the problems in aligning quantitative findings with existing classifications or findings from qualitative approaches extend beyond the specific realm of bibliometric research.

The recent findings reported above contradict the prior experiences of the authors, who, as analysts working for a commercial bibliometric service provider, have been able to deploy XDR indicators in the formal program evaluations of a number of funding initiatives. The authors have found two bibliometric indicators, disciplinary diversity in authorships (DDA) and disciplinary diversity in references (DDR), to be of value in characterizing the scientific output of funding programs and hypothesizing on their likely mechanisms for steering research practices.

The authors note that there is a clear expansion in the number of currently operating funding programs seeking out measurements on the degree to which supported projects have achieved the XDR outcomes expected from the funding mechanism. On this basis, this paper is motivated by a pragmatic aim to support the managers of a large number of existing programs (Gleed & Marchant, 2016; Rylance, 2015) and other policymakers in developing capacity for the critical assessment of their XDR initiatives. We cannot touch here on the question of how the expansion of XDR funding programs impacts the scientific system more broadly; or if, when and how it is desirable to deploy XDR research funding – although the authors have recently reported that XDR research may be associated with greater odds of policy-relevant outcomes than monodisciplinary research (Pinheiro, Vignola-Gagné, & Campbell, 2021).

In this paper, we aim to test whether our own observations can be generalized by applying these measures across a large number of funding programs, expected to account for different mechanisms to support (or not) XDR; from purely excellence-driven investigator-level grants to initiatives that support highly diverse, transdisciplinary collaborations. We argue that breaking down publication sets by funding programs is currently the best possible approach in trying to evaluate the robustness of XDR indicators. We contend that in no other level of analysis (individual researcher; department; institution; networks, to take just a few example) can there be such a sustained orientation towards XDR practice as that which can be achieved in a dedicated funding program. Funding programs are also a major, and quite possibly the most important socio-organizational mechanism through which XDR is being promoted towards and adopted by researchers. Additionally, developing approaches for measuring XDR outcomes of funding programs is of practical relevance in the context of applied bibliometrics.

We provide a preliminary test of the usefulness and validity of XDR indicators for evaluating the XDR-fostering potential of funding programs. We draw on descriptive statistics, comparing expected XDR practice levels based on qualitative appraisal of program design characteristics and instruments, against measurements obtained through our two XDR indicators. A starting hypothesis here is that funding is the most likely and possibly effective mean through which researchers are exhorted to engage in more XDR research. While some researchers certainly come to engage in XDR practices as part of their intellectual or networking trajectories, our assumption is that funding provides the logistic basis for acting on this ambition.

The methods section will detail the indicators of interdisciplinarity and multidisciplinarity used, with more extensive detail for the second as it amounts to a novel variation of the Rao-Stirling computation applied to publication authorship characteristics. The method section will also include a brief description of the funding programs selected for inclusion in this exercise. Findings are presented at the level of the individual funding programs retained, for DDA, DDR and other relevant dimensions. Findings from follow-up investigations are also presented to test specifically the expectation that various groupings of publications (e.g., by research area) could achieve exhibit intellectual integration (DDR) independently of team-based collaboration (DDA). In this paper, this was done by measuring XDR for Scopus topics of prominence. We conclude with a discussion of these findings and next steps.

Methods

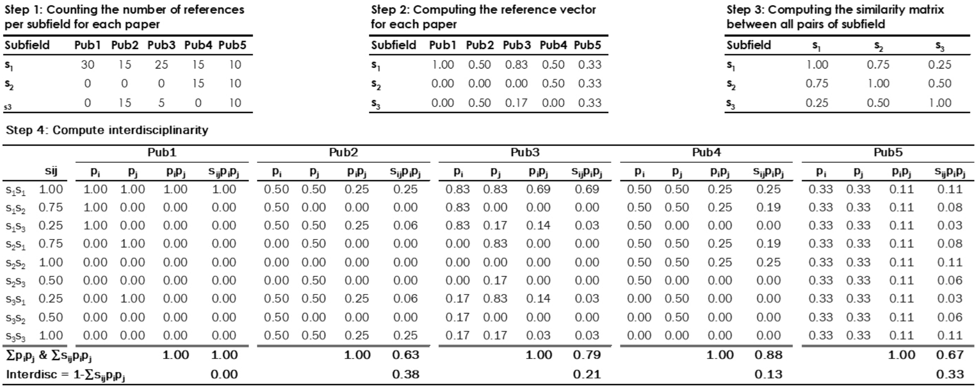
XDR indicators

Cross-disciplinarity at the paper level was captured through two lenses (see also Pinheiro et al., 2021): disciplinary diversity of cited references (DDR; tracks diversity of integrated knowledge) and disciplinary diversity of contributing authors (DDA). The former is equivalent to the integration metrics of Porter and Rafols (2009), relying on a hybrid (journal- and article-based) version of the Science-Metrix classification of science to classify a paper's cited references by subfield (see below for more details) (Rivest, Vignola-Gagné, & Archambault, 2021). Using their method, each paper in Scopus is assigned a DDR score from 0 (i.e., completely following predominant citation patterns) to 1, the latter being extremely interdisciplinary (i.e., diverging completely from normal citation patterns, integrating knowledge from areas that others do not). Per their method, a paper's DDR score is computed as:

$$Interdisciplinarity = 1 - \sum_{i,j} s_{ij} p_i p_j$$

Where p_i and p_j are the respective proportions of references in subfields i and j in a paper's reference list. The summation is taken over all cells of the subfield-by-subfield similarity matrix, accounting for all subfields in Science-Metrix taxonomy. s_{ij} is the similarity between subfields i and j and captures how close (or distant, by taking 1-sij as in the above formula) the integrated subfields are in a given paper. The similarity matrix between subfields was obtained by computing the cosine similarity between each pair of subfields, each subfield being represented by an ordered vector of 174 dimensions (one per Science-Metrix subfield). The values populating a given subfield vector represent the proportion of instances in which it was co-cited with each of the 174 subfields in the whole of Scopus. Table I depicts the computation for five papers in a system with three subfields (e.g., s1 = forestry, s2 = applied mathematics, and s3 = optics; the similarities in the example are fictive).

Table I Illustration of the computation of a paper's interdisciplinarity/DDR score



Step 1: Counting the number of references per subfield for each paper

Subfield	Pub1	Pub2	Pub3	Pub4	Pub5
s_1	30	15	25	15	10
s_2	0	0	0	15	10
s3	0	15	5	0	10

Step 2: Computing the reference vector for each paper

Subfield	Pub1	Pub2	Pub3	Pub4	Pub5
s_1	1.00	0.50	0.83	0.50	0.33
s_2	0.00	0.00	0.00	0.50	0.33
s_3	0.00	0.50	0.17	0.00	0.33

Step 3: Computing the similarity matrix between all pairs of subfield

Subfield	s_1	s_2	s_3
s_1	1.00	0.75	0.25
s_2	0.75	1.00	0.50
s_3	0.25	0.50	1.00

Step 4: Compute interdisciplinarity

		Pub1				Pub2				Pub3				Pub4				Pub5			
	sij	p_i	p_j	p_ip_j	$s_{ij}p_ip_j$	p_i	p_j	p_ip_j	$s_{ij}p_ip_j$	p_i	p_j	p_ip_j	$s_{ij}p_ip_j$	p_i	p_j	p_ip_j	$s_{ij}p_ip_j$	p_i	p_j	p_ip_j	$s_{ij}p_ip_j$
s_1s_1	1.00	1.00	1.00	1.00	1.00	0.50	0.50	0.25	0.25	0.83	0.83	0.69	0.69	0.50	0.50	0.25	0.25	0.33	0.33	0.11	0.11
s_1s_2	0.75	1.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.83	0.00	0.00	0.00	0.50	0.50	0.25	0.19	0.33	0.33	0.11	0.08
s_1s_3	0.25	1.00	0.00	0.00	0.00	0.50	0.50	0.25	0.06	0.83	0.17	0.14	0.03	0.50	0.00	0.00	0.00	0.33	0.33	0.11	0.03
s_2s_1	0.75	0.00	1.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.83	0.00	0.00	0.50	0.50	0.25	0.19	0.33	0.33	0.11	0.08
s_2s_2	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.25	0.25	0.33	0.33	0.11	0.11
s_2s_3	0.50	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.17	0.00	0.00	0.50	0.00	0.00	0.00	0.33	0.33	0.11	0.06
s_3s_1	0.25	0.00	1.00	0.00	0.00	0.50	0.50	0.25	0.06	0.17	0.83	0.14	0.03	0.00	0.50	0.00	0.00	0.33	0.33	0.11	0.03
s_3s_2	0.50	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.33	0.33	0.11	0.06
s_3s_3	1.00	0.00	0.00	0.00	0.00	0.50	0.50	0.25	0.25	0.17	0.17	0.03	0.03	0.00	0.00	0.00	0.00	0.33	0.33	0.11	0.11
Σp_ip_j & $\Sigma s_{ij}p_ip_j$				1.00	1.00			1.00	0.63			1.00	0.79			1.00	0.88			1.00	0.67
Interdisc = $1-\Sigma s_{ij}p_ip_j$					0.00				0.38				0.21				0.13				0.33

Per this example, one can see that subfields 1 and 3 are the least similar (or most dissimilar), and that subfield 2 is most similar to subfield 1, while at the same time being more similar to subfield 3 than subfield 1 is to subfield 3. Thus, the combination of subfields 1 and 3 should contribute most to DDR, followed by subfields 2 and 3, and then subfields 1 and 2. Of course, if only one subfield is cited in a paper (either subfield 1, 2 or 3) as in publication #1, DDR will equal 0 (monodisciplinary paper). Publication #2 strictly cited the two most distant subfields (i.e., 1 and 3) in equal proportions. This resulted in the highest DDR score (0.38) across the five publications shown in the example. Publication #3 also strictly cited subfield 1 and 3, but in very unequal proportions. Accordingly, its DDR score (0.21) is lower than for publication #2 (0.38). Publication #4 strictly cited subfields 1 and 2 in equal proportions. Since they are the least distant (or most similar) subfields, its score (0.13) is lower than for publication #2 (0.38) even though they both have their references uniformly distributed across cited subfields. Finally, publication #5 is the only one to have cited all three subfields, and this in equal proportions. Due to the balanced integration of a greater variety of disciplines, it scores higher than publication #3 and #4, but not more than publication #2. This is because although publication #2 only cited two subfields instead of three, the average distance between the integrated disciplines in publication #2 is higher than in publication #5—that is, the additional subfield cited in publication #5 reduces the overall "intellectual" distance in the pool of referenced work.

In computing DDR, the proximity between integrated subfields is computed as the cosine similarity matrix between subfields, relying on the subfield co-citation network in the whole of Scopus (using peer-reviewed publications; i.e., journal articles, conference papers and reviews). The classification used to categorise publications by subfield can have an impact on the resulting DDR scores. It is important for such a classification to offer enough granularity to enable the detection of relevant disciplinary mixes. For example, carrying out the analysis at the level of large scientific domains (e.g., natural sciences, engineering, health sciences, social sciences and humanities) would disregard significant differences in scientific culture, methods and tools within each of those domains (e.g., between biology, chemistry and physics within the natural sciences). At the other end of the spectrum, too much granularity would introduce noise in the analysis, detracting our attention from those disciplinary mixes that matter most. For example, in most cases of funders promoting cross-disciplinary work, it would appear irrelevant to know that an entomologist working on species X collaborated with an entomologist working on species Y, but it would matter to know that an entomologist working on the ecology of species X collaborated with a geneticist to study the population genetics of species X. Selecting the appropriate classification to study DDR is not a trivial choice. Ideally, one wants to pick a structure that somewhat reflects the current division of staff in academic departments at higher education institutions. One such classification was developed by Science-Metrix (Archambault et al., 2011) and was recently recognised as the most accurate journal-based classification of scientific papers (Klavans & Boyack, 2017).

In its most recent version, generalist journals publishing papers in diverse fields (e.g., Nature, Science, PNAS, PLOS One) were reclassified in the most appropriate category at the paper level using a machine learning algorithm. The updated classification tree now includes 174 subfields distributed across 20 fields and 5 domains. All 174 subfields are used in computing DDR.

Since the similarity matrix was built using all years in Scopus (since 1996) irrespective of a paper's publication year, recent papers may naturally have a greater tendency towards higher DDR scores than older ones. What was a rare disciplinary connection in 1996 may now be quite common. For this reason, DDR was normalized by publication year to uncover, for each year, those papers that stand out on DDR. Additionally, subfields are not all equally covered in Scopus, meaning that the proportion of a paper's references that are unclassified, and thus overlooked in computing DDR, varies across them. To eliminate the effects of these coverage biases, DDR was also normalized by subfield. As a result, the emphasis was placed on those publications that stand out most on the DDR scale within each subfield and year.

The latter XDR indicator, DDA, measured diversity as reflected in the prior disciplinary background of a paper's co-authors (team multidisciplinarity). Authors were disambiguated using Scopus author IDs, which produce reliable results at scale (Campbell & Struck, 2019). Science-Metrix subfields were assigned to authors on a given paper based on their prior publications enabling the disciplinary background of authors to evolve through time. Percentage shares of prior publications by Science-Metrix subfield were integrated in a combined vector (see also (Zuo & Zhao, 2018) for a related approach).

Some of the limitations linked to inferring the disciplinary profiles of authors using this approach can already be highlighted. First, a large portion of researchers (e.g., students, some of which later left academia) published very few Scopus-indexed publications (or even only one publication). For these researchers, the assignation of a subfield vector is based on scant data. Thus, the resulting vector may not be representative of their “true” disciplinary background. As an example, researchers with few prior publications in Scopus may be assigned subfields other than their own. This could happen if they previously provided methodological support to the research teams of their previous publications (e.g., a graduate student in computer sciences providing support to biology papers). In other cases, graduate students could have had no previous publications, in which case they would not contribute to the score even though they effectively contributed to the paper. Despite these limitations, we observed a positive correlation between normalized DDA and DDR whose moderate strength (Pearson coefficient = 0.38) is in line with our expectation. While DDA should be conducive to DDR, it may not always lead to high DDR and DDR can in theory result from the work of a single researcher.

Compared to DDR, where each of a paper’s references are assigned a unique subfield, authors can be assigned multiple subfields in computing a paper’s DDA score. This is a key difference requiring adaptation to the Rao-Stirling diversity index. Once all authors on a paper are assigned a subfield vector, a subfield vector for the paper is created in two steps (see Table 2 for examples in a 3-subfield system). First, the authors’ subfield vectors are averaged together. Then, the averaged vector is subtracted from the authors’ subfield vectors, and the absolute differences are kept. Conceptually, what remains are the distances between the authors’ disciplinary profiles and the average disciplinary profile of the publication. If all authors share the exact same profile (i.e., no between-author variation in disciplinary profiles), and this even if within-author variation is present as for publication #1 in Table 2, differences will be null, leaving all authors’ subfield vectors empty, which is to say that the paper will score 0 on the DDA scale (i.e., to be interpreted as monodisciplinary). Otherwise, the differences will, once averaged across authors, reflect the paper’s disciplinary diversity stemming from the within- and between-author disciplinary diversity (see publication #2 in Table 2).

In measuring DDA, emphasis is placed on the collaborative aspect, which is the common denominator of most funding programs promoting cross-disciplinary research. Accordingly, within-author diversity (an author active in more than one subfield) will not contribute to the paper’s DDA score if there is no difference across authors' disciplinary profiles or if there is only one author (all single-author papers will score 0). It is important to recall that a single-author paper, even if authored by a researcher who previously published in diverse subfields, even distant ones, may not itself have resulted from the integration of knowledge from different subfields. A diverse publication profile for an author may simply result from him or her having changed discipline while approaching research in his or her new field in a traditional way (i.e., adopting the research practices of this new field in a monodisciplinary way). The DDA indicator aims to capture the disciplinary diversity across the co-authors of a paper, not the disciplinary diversity of an author’s publication profile. If a single-author paper is produced by a researcher with a diversified publication profile, the odds of the researcher integrating knowledge from different disciplines in producing this paper is probably higher than for a researcher with a “narrow” publication profile. However, this should be captured through DDR, not DDA.

The above approach for producing a paper’s subfield vector using author information runs the risk of counting as monodisciplinary all papers produced by, for example, two authors, a biologist and a physicist, who always published together, resulting in the exact same subfield vector for each of them (e.g., 50% in biology and 50% in physics). This, even though they truly provided distinct disciplinary inputs to their common work. This is a key limitation of the proposed metric. In practice, however, such cases are very unlikely. Given the number of subfields in the classification used (i.e., 174), most authors will publish in many subfields over their career (so more senior authors are likely to have more variation in their subfield vectors) making it unlikely for any given paper’s authors to share exactly the same subfield vector.

To compute the DDA of a given paper, the Rao-Stirling index as depicted above for DDR is applied to its resulting subfield vector obtained using the above procedure with author disciplinary profiles. Briefly, a paper’s reference vector (obtained in step 2 for DDR) is substituted by its paper’s subfield vector. Contrary to a paper’s reference vector, note that the sum of values across subfields for a paper’s subfield vector based on authors may add up to more or less than 1 (see publication #2 in Table 2). This implies that the adapted Rao-Stirling index for DDA is not bounded between 0 and 1 as for DDR. Instead, it ranges from 0 (in theory) to infinity (in practice, it rarely exceeds 1).

To ensure that the disciplinary profiles of a paper’s authors are adequately captured, DDA is only computed for papers with at least two effective authors (i.e., authors with an assigned subfield vector). For a score to be computed, the effective number of authors must also represent at least 20% of the real number of authors. A DDA score is not computed for any paper not matching these two conditions. Single-author publications are by default monodisciplinary papers (see above explanations).

Once a paper’s DDA score is obtained, it is normalized (by year and subfield) in the same way as DDR (see above).

Table 2Illustration of the computation of a paper’s subfield vector based on its authors’ disciplinary profiles

Step 1: Computing the average disciplinary profile of the paper

	Publication #1				Publication #2			
	Author subfield vector			Paper's avg subfield vector	Author subfield vector			Paper's avg subfield vector
	A1	A2	A3		A1	A2	A3	
S1	0.50	0.50	0.50	=(0.50+0.50+0.50)/3 = 0.50	0.50	0.50	0.00	0.33
S2	0.50	0.50	0.50	0.50	0.50	0.50	0.00	0.33
S3	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.33

Step 2: Computing the paper's subfield vector from the differences between author subfield vectors and the paper's avg subfield vector

	Publication #1				Publication #2			
	Difference (Author subfield vector - paper's avg subfield vector)			Paper's subfield vector	Difference (Author subfield vector - paper's avg subfield vector)			Paper's subfield vector
	A1	A2	A3		A1	A2	A3	
S1	=ABS(0.50-0.50)=0.00	0.00	0.00	0.00	0.17	0.17	0.33	0.22
S2	0.00	0.00	0.00	0.00	0.17	0.17	0.33	0.22
S3	0.00	0.00	0.00	0.00	0.33	0.33	0.67	0.44

DDR and DDA are both evaluated in two forms: with indices based on averages of normalized scores and, by calculating shares of publications amongst the 10% most interdisciplinary or multidisciplinary within their subfield.

Selection of funding initiatives

The authors regularly lead bibliometric components of funding program evaluations for funder globally. This experience provides them with either direct or indirect insight into the administration of a number of funding programs worldwide. Additionally, the authors have also conducted desk research on multiple additional programs as part of comparator selection approaches. Selection of funding programs to be included as analytical groups in this analysis drew from this prior experience. Prior publications on the implementation of some of these programs also provided further support in characterizing their degree of XDR focus.

Funding programs to be included in the analysis were selected with the expectation that they represent different cases of funding mechanisms:

- “Generic” funding programs oriented towards **investigator-driven research** (expected to display low or average scores for both DDR and DDA): the Canadian NSERC Discovery program; the Australian ARC Discovery program; the NSFC National Program for Basic Research; and two “investigator-initiated research projects” programs from the US NSF Division of Molecular and Cellular Biosciences, on cellular dynamics and function (MCB-CDF); and on genetic mechanisms (MCB-GM).
- Funding programs supporting above all **individual scientific excellence** (expected to display low or average scores for DDA; and no fixed expectation for DDR): the European Research Council; the Fulbright Program (Senior Scholar program); the MacArthur Foundation (Fellowships and so-called “genius grants”).
- Funding programs supporting **individual XDR projects** (expected to display high DDR but low or average DDA): the Human Frontiers Science Program (HFSP) Cross-Disciplinarity Fellowships (CDF) competition (Science-Metrix, 2018); the NSF MCB Investigator-initiated research projects – systems and synthetic biology (MCB-SSB); and on molecular biophysics (MCB-MBP).
- Funding programs supporting **collaborative XDR consortia**, without an explicitly stated goal to support intellectual integration (expected to display high DDA, and low or average DDR): the SNFC Fund for Creative Research Groups; and the US NSF Research Coordination Networks (Garner, Porter, Newman, & Crowl, 2012).
- Funding programs supporting **transformative, transdisciplinary or highly integrative research conducted by diverse consortia** (expected to display both high DDA and DDR): the National Academies Keck Futures Initiative (NAKFI); Swiss SNSF National Centres of Competence in Research (NCCR) (Schneider, Buser, Keller, Tribaldos, & Rist, 2019), Sinergia and National Research Program (NRP) programs; US NSF Science and Technology Centers; and Synthesis Centers (Hackett et al 2021); and the Australian ARC Centres of Excellence (COE) initiative.

Constituting of publication sets by funding program

Three main strategies were used to assemble sets of publications associated to the funding programs under analysis:

- Publication sets were available from prior evaluation work and were initially assembled from a funder-curated author list (HFSP CDF program)
- Publication sets were constituted specifically for this analysis by retrieving curated publication lists from a public database (some NSF programs)
- Publication sets were constituted specifically for this analysis by querying Scopus acknowledgement records (remaining programs)

In matching publications to funding program bins, an important assumption made was that most papers (and indeed even many individual authors) are supported by multiple funding programs, concurrently. We expect most researchers with XDR funding to simultaneously hold funding from investigator-driven, monodisciplinary programs. In the aggregate, scores for papers funded by XDR programs should tend towards higher DDR/DDA scores, despite cases of single papers in the overall sample having contributed observations to both XDR and non-XDR programs.

Precision of the delineation of publication sets by querying Scopus funding acknowledgements was evaluated by randomly selecting 300 within the overall pool of publications retrieved in this manner and manually validating their match to funding programs of interest. Precision in this sample was measured at 99.0%.

Complementary search for funding programs

Initial results did not conform to expectations, especially in the absence of findings showing that funding programs can foster DDR without also fostering DDA. To demonstrate that it is theoretically possible to achieve high DDR in the absence of high DDA, DDA and DDR measurements were also taken for Scopus "topics of prominence" (Klavans & Boyack, 2017). Capturing instances of high DDR/low DDA for some topics would showcase that intellectual integration of diverse bodies of knowledge is possible outside the context of a partnership between researchers of different disciplines. In that case, the unexpected results for XDR programs could be explained by the simple fact that one of the easiest mechanisms to foster DDR is most likely through collaborative work spanning many disciplines (i.e., DDA).

Bootstrapping to calculate stability intervals for DDR and DDA measurements

Bootstrapping was used to produce stability intervals for the average DDR and DDA measurements for each group. A total of 1,999 re-samples were run to produce point estimates (at the 50th percentile of observations' distribution); lower boundary of the stability interval (measured at the 2.5th percentile) and upper boundary (recorded at the 97.5th percentile).

Results

Main benchmarking *of funding programs*

Most findings presented in Table 3 fell into two well defined groups: either non-XDR programs (first two blocks of programs in Table 3) which generally (5 out of 8 or 63%) did not foster XDR publications (DDA and DDR both between 0.95 and 1.05); and XDR programs that most often (9 out of 12 or 75%) fostered DDA, and generally also DDR, but often with a more notable effect on the DDA dimension (keeping in mind that variance on the DDA scale is generally greater than on the DDR scale).

Patterns in shares of publications falling in the 10% most interdisciplinary (DDR 10%) and multidisciplinary (DDA 10%) papers generally, but not always, followed those based on average DDR (simply DDR in Table 1) and DDA.

Four standard investigator-driven grant programs conformed to expectation: they displayed average scores for both DDA and DDR indices, or even lower (MCB-GM). The MCB-CDF program displayed a moderately high DDA score however (1.15).

From these results, we observe that high DDA can occur without high DDR (e.g., MCB-CDF, SNSF-Sinergia), which is in line with our interpretation of both phenomena (see particularly Figure 1). However, most of the time, high DDA is associated with high DDR, which seems logical as DDA would, on average, appear conducive to DDR. From a funder's perspective, high DDR did not come up in the absence of high DDA. This could be explained by the difficulty (most notably for funders) of promoting high DDR through a mechanism other than collaboration in teams of researchers from diverse disciplinary background (high DDA).

Table 3 – Comparison between expected and measured XDR levels for selected Funding programs

<table>
<tr><th>Country</th><th>Agency</th><th>Program</th><th>Count</th><th>DDR</th><th>DDR
10%</th><th>DDA</th><th>DDA
10%</th><th>avg
refs</th><th>avg
auth</th><th>med
auth</th></tr>
<tr><td colspan="11">**Standard investigator-driven grants (expected average or low DDR and DDA)**</td></tr>
<tr><td>Can</td><td>NSERC</td><td>Discovery</td><td>34,616</td><td>**1.00**
0.997|-|1.006</td><td>9.7%
9.359|-| 9.975</td><td>**0.95**
0.943|-|0.964</td><td>9.0%
8.702|-| 9.335</td><td>52</td><td>4</td><td>3</td></tr>
<tr><td>Aus</td><td>ARC</td><td>Discovery</td><td>21,128</td><td>**1.00**
0.991|-|1.002</td><td>9.3%
8.909|-| 9.702</td><td>**1.00**
0.983|-|1.009</td><td>9.2%
8.772|-| 9.630</td><td>51</td><td>5</td><td>4</td></tr>
<tr><td>USA</td><td>NSF</td><td>MCB-CDF</td><td>5,217</td><td>**1.02**
1.012|-|1.028</td><td>9.7%
8.857|-|10.525</td><td>**1.15**
1.124|-|1.183</td><td>14.2%
13.199|-|15.251</td><td>56</td><td>5</td><td>4</td></tr>
<tr><td>USA</td><td>NSF</td><td>MCB-GM</td><td>5,172</td><td>**0.94**
0.936|-|0.952</td><td>4.3%
3.698|-| 4.848</td><td>**0.92**
0.900|-|0.951</td><td>8.7%
7.899|-| 9.554</td><td>56</td><td>6</td><td>4</td></tr>
<tr><td>China</td><td>NSFC</td><td>Basic Research (973)</td><td>85,596</td><td>**1.00**
1.000|-|1.005</td><td>8.4%
8.227|-| 8.580</td><td>**1.00**
0.994|-|1.006</td><td>8.8%
8.653|-| 9.041</td><td>39</td><td>6</td><td>5</td></tr>
<tr><td colspan="11">**Individual scientific excellence grants (expected average or low DDR and DDA)**</td></tr>
<tr><td>USA</td><td>Fulbright Found.</td><td>Senior Scholar</td><td>286</td><td>**1.06**
1.010|-|1.116</td><td>11.4%
7.478|-|15.487</td><td>**1.11**
0.980|-|1.243</td><td>10.8%
6.604|-|15.000</td><td>51</td><td>4</td><td>3</td></tr>
<tr><td>USA</td><td>MacArthur Found.</td><td>Fellowships and genius grants</td><td>79</td><td>**1.24**
1.120|-|1.349</td><td>29.8%
17.544|-|42.105</td><td>**1.40**
1.091|-|1.735</td><td>23.2%
12.500|-|33.929</td><td>58</td><td>6</td><td>3</td></tr>
<tr><td>EU</td><td>ERC</td><td>All</td><td>66,571</td><td>**0.98**
0.978|-|0.984</td><td>8.4%
8.206|-| 8.616</td><td>**0.97**
0.960|-|0.975</td><td>9.4%
9.180|-| 9.643</td><td>57</td><td>20</td><td>5</td></tr>
<tr><td colspan="11">**Grants to individual XDR projects (expected high DDR and average or low DDA)**</td></tr>
<tr><td>Intl.</td><td>HFSP</td><td>CDF</td><td>266</td><td>**1.13**
1.083|-|1.164</td><td>21.3%
16.297|-|26.253</td><td>**1.83**
1.695|-|1.971</td><td>34.6%
28.794|-|40.467</td><td>49</td><td>5</td><td>4</td></tr>
<tr><td>USA</td><td>NSF</td><td>MCB-SSB</td><td>3,937</td><td>**1.07**
1.061|-|1.077</td><td>10.9%
9.951|-|11.910</td><td>**1.30**
1.269|-|1.338</td><td>17.4%
16.158|-|18.641</td><td>56</td><td>6</td><td>4</td></tr>
<tr><td>USA</td><td>NSF</td><td>MCB-MBP</td><td>10,720</td><td>**1.08**
1.072|-|1.083</td><td>10.2%
9.665|-|10.826</td><td>**1.22**
1.201|-|1.240</td><td>13.9%
13.168|-|14.554</td><td>53</td><td>5</td><td>4</td></tr>
<tr><td colspan="11">**Collaborative XDR consortia (expected average or low DDR and high DDA)**</td></tr>
<tr><td>China</td><td>NSFC</td><td>Creative Res. Groups</td><td>3,508</td><td>**1.04**
1.028|-|1.053</td><td>12.1%
11.072|-|13.248</td><td>**1.02**
0.991|-|1.049</td><td>9.1%
8.103|-|10.062</td><td>37</td><td>5</td><td>5</td></tr>
<tr><td>USA</td><td>NSF</td><td>RCN</td><td>535</td><td>**1.15**
1.120|-|1.180</td><td>18.1%
14.673|-|21.378</td><td>**1.50**
1.399|-|1.606</td><td>23.3%
19.706|-|27.044</td><td>60</td><td>6</td><td>5</td></tr>
<tr><td colspan="11">**Transformative, transdisciplinary or highly integrative research by diverse consortia (expected high DDR and DDA)**</td></tr>
<tr><td>USA</td><td>Nat. Academies and Keck Found.</td><td>NAFKI</td><td>128</td><td>**1.22**
1.156|-|1.278</td><td>32.6%
24.690|-|40.639</td><td>**2.12**
1.803|-|2.425</td><td>38.9%
30.088|-|48.673</td><td>52</td><td>4</td><td>4</td></tr>
<tr><td>USA</td><td>NSF</td><td>Synthesis Centers</td><td>1,198</td><td>**1.11**
1.095|-|1.132</td><td>19.7%
17.467|-|21.994</td><td>**1.36**
1.297|-|1.431</td><td>18.5%
16.276|-|20.953</td><td>65</td><td>6</td><td>4</td></tr>
<tr><td>USA</td><td>NSF</td><td>STC</td><td>497</td><td>**1.13**
1.088|-|1.182</td><td>14.4%
11.260|-|17.685</td><td>**1.09**
0.998|-|1.195</td><td>10.5%
7.912|-|13.407</td><td>46</td><td>5</td><td>4</td></tr>
<tr><td>Aus</td><td>ARC</td><td>CoE</td><td>5,761</td><td>**0.95**
0.941|-|0.963</td><td>9.0%
8.250|-| 9.699</td><td>**0.96**
0.938|-|0.990</td><td>9.5%
8.782|-|10.313</td><td>59</td><td>8</td><td>5</td></tr>
<tr><td>CH</td><td>SNSF</td><td>NCCR</td><td>3,943</td><td>**0.99**
0.982|-|1.005</td><td>7.5%
6.688|-| 8.262</td><td>**0.98**
0.949|-|1.008</td><td>8.6%
7.697|-| 9.483</td><td>53</td><td>6</td><td>5</td></tr>
<tr><td>CH</td><td>SNSF</td><td>NRP</td><td>942</td><td>**1.10**
1.075|-|1.121</td><td>16.4%
14.158|-|18.775</td><td>**1.33**
1.268|-|1.403</td><td>18.0%
15.653|-|20.383</td><td>53</td><td>6</td><td>5</td></tr>
<tr><td>CH</td><td>SNSF</td><td>Sinergia</td><td>925</td><td>**1.00**
0.979|-|1.018</td><td>7.7%
6.105|-| 9.376</td><td>**1.11**
1.044|-|1.167</td><td>11.5%
9.368|-|13.680</td><td>58</td><td>8</td><td>6</td></tr>
</table>

Follow-up identification of DDR-oriented topics of prominence

As an additional step to illustrate the independence between DDR and DDA (we saw in the methods that there is only a moderate positive correlation between them at the paper level in the full database, Pearson r = 0.38), Scopus topics of prominence were investigated with the specific goal of identifying topics with a higher DDR score than DDA score. That is, the intention was to find specialties where investigators and/or research teams tend to display average or low levels of disciplinary diversity but still integrate intellectual insights from a diversity of research fields. This design was operationalized as identifying all topics of prominence with an average DDR score 0.25 points above their average DDA score.

This stream of measurements identified more than 14,500 topics of prominence (from a total of more than 97,000 such topics available in Scopus) falling within the criteria specified above. Of these topics, 2,087 were associated with 1,000 publications or more. Table 4 presents the top 10 (on publication volume) topics of prominence retrieved in this step.

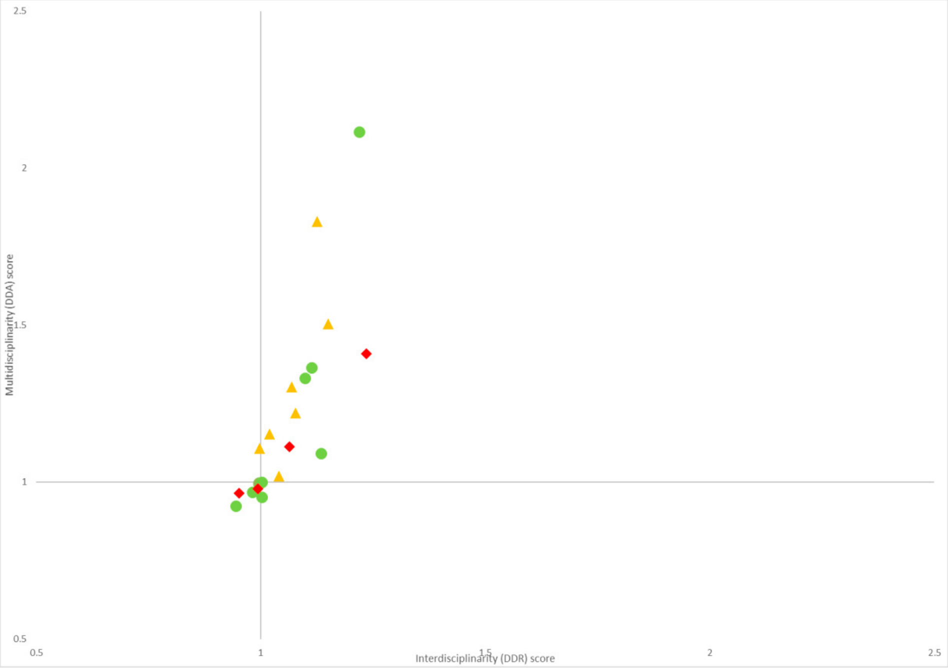


Figure 1. Distribution of 20 funding programs for their support of DDR and DDA research practices. Point shape denotes conformation to expectations. Circles denote funding programs conforming to expectations; triangles, programs deviating from expectations on one dimension; diamonds, programs deviating from expectations on both dimensions of interest.

Table 4 – Instances of Scopus topics of prominence with higher DDR than DDA

Topic of prominence	Count	DDR	DDR 10%	DDA	DDA 10%	Avg Refs	Avg auth	Med auth
Superconductors (materials) ; Superconductivity ; Iron-based superconductor	8,021	1.07	1.0%	0.60	1.9%	37	6	6
Cloud computing ; Clouds ; Machine placement	7,364	1.28	8.3%	0.98	7.5%	26	3	3
Activation analysis ; Catalysis ; Bond functionalization	7,220	0.59	0.0%	0.33	1.2%	69	4	4
Dark energy ; Models ; Holographic dark	6,544	1.53	11.1%	0.91	7.4%	54	2	2
Angular momentum ; Vortex flow ; Carrying orbital	6,256	0.96	2.9%	0.61	3.1%	22	4	4
Suburethral Slings ; Urinary Incontinence, Stress ; Subjective cure	6,200	0.88	0.4%	0.57	2.9%	18	3	3
Gamma ray bursts ; Afterglows ; Afterglow emission	6,071	1.15	4.3%	0.74	4.5%	44	7	3
Metamaterials ; Resonators ; Left-handed metamaterial	5,992	1.42	16.2%	1.07	10.3%	19	3	3
Nonexpansive mapping ; Strong convergence ; Pseudocontractive mappings	5,916	1.45	11.5%	0.82	5.8%	23	1	2
Multiple Myeloma ; Patients ; Diagnosed multiple	5,884	0.87	0.4%	0.56	2.6%	31	6	4

These findings at the topics of prominence aggregation-level provide some confirmation that DDR has some degree of independence from DDA in circumstances where individuals or monodisciplinary teams draw knowledge from a range of disciplines. Many of the topical publication sets in Table 2 display large gaps in scores between DDR and DDA, with DDR often being above expectation while DDA is below expectation. For instance, DDR scores were also often higher than previously observed at the funding program level (including scores ranging from 1.28 and 1.53). Dark energy physics (DDR of 1.53) may not be associated a priori with high interdisciplinarity, but it should be kept in mind that the associated publications' scores have been normalized against average levels of intellectual integration in the particular Science-Metrix subfields in which publications in this topic fell. These results, combined with those above for funding programs, support the idea that funders best leverage point to fuel DDR is likely through multidisciplinary collaboration (or DDA).

Limitations and discussion

Overall, the DDR and DDA indicators in their current specification appeared to successfully differentiate between 1) traditional grants for "investigator-driven" research, on the one hand (non-XDR programs); and 2) most competition for mission-oriented grants that included at least one component to support XDR practices (XDR programs), although these programs tended to score high on both dimensions, irrelevant of formal program design focus towards interdisciplinary, multidisciplinarity or a combination of both.

The mismatches between expectations and measurements are likely to be explained by one or a combination of multiple factors such as: XDR programs being more or less performing than one might expect from the mechanism(s) they implemented to fuel XDR, and the challenges of qualitatively gauging a program's expected level of XDR. In retrospect, and following additional desk research, this latter factor appears to have played a role. In some cases, the initial characterization of funding programs in terms of XDR clearly failed to reveal their cross-disciplinarity character. This was exemplified by the NSF program on Molecular Biophysics, which despite being initially described as "investigator-driven" on the program's website, was noted to include a "collaborative" grants component when extraction from the NSD award database was conducted. In the case of the HFSP CDF program, which one of the authors has evaluated, disaggregating the sub-dimensions of DDA and DDR (as done in Hackett et al., 2021) might have better captured the program's specificities, with an operationalization of DDR centered around balance and distance but not variety.

Some of the transdisciplinary programs included here aimed to produce non-academic outcomes as much as peer-reviewed publications. This observation raises the possibility that projects supported by XDR programs may produce rather conventional articles while simultaneously engaging in sustained relationships with local stakeholders and public engagement, to take just an hypothetical example.

It must also be considered that the design and formulation of policy instruments for fostering XDR will not always lead to successful implementation of these means. The SNSF-NCCR appeared to the authors as one of the programs with the strongest XDR designs in our initial desk research, but the article by Schneider et al also show the range of obstacles facing the successful implementation of such a program, and most notably, to ensure the participation of researchers in XDR practices. In can also be considered that prior evaluations by the authors of individual consortia within the SNSF NCCR program and within the ARC CoE program had shown individually selected projects with a clear focus towards XDR to have indeed achieved high scores on both the DDR and DDA dimensions.

It must be noted that "discovery" or excellence funding programs tend to be designed and implemented quite differently from the typical funding program explicitly aiming to support XDR. "Non-XDR funding programs" tend to be recurring, foundational granting programs that provide the largest shares of researchers within a given scientific system; XDR funding programs tend to be ad hoc, mission-oriented instruments with a narrower audience. It is currently unclear how these differing program design features may play into the findings recorded here.

The next steps for this research project are to input some of the findings presented here in a regression analysis that will be better control for confounding factors and help specify the relationship between DDR and DDA. To do so, the authors will also attempt to correlated DDR and DDA measurements with those obtained from other indicators such as "Div"(Leydesdorff, Wagner, & Bornmann, 2019) or the disaggregated components of DDR and DDA, as proposed by (Hackett et al., 2021; J. Wang, Thijs, & Glänzel, 2015).

References

Campbell, D., & Struck, B. (2019). Reliability of Scopus author identifiers (AUIDs) for research evaluation purposes at different scales. *17th International Conference of the International Society for Scientometrics and Informetrics (ISSI), 2–5 September 2019, Proceedings Vol. II*, 1276–1287. Retrieved from http://issi-society.org/publications/issi-conference-proceedings/proceedings-of-issi-2019/

Garner, J. G., Porter, A. L., Newman, N. C., & Crowl, T. A. (2012). Assessing research network and disciplinary engagement changes induced by an NSF program. *Research Evaluation*. https://doi.org/10.1093/reseval/rvs004

Gleed, A., & Marchant, D. (2016). *Interdisciplinarity survey report for the Global Research Council 2016 annual meeting*. Stockport.

Hackett, E. J., Leahey, E., Parker, J. N., Rafols, I., Hampton, S. E., Corte, U., … Vision, T. J. (2021). Do synthesis centers synthesize? A semantic analysis of topical diversity in research. *Research Policy*, *50*(1). https://doi.org/10.1016/j.respol.2020.104069

Klavans, R., & Boyack, K. W. (2017). Research portfolio analysis and topic prominence. *Journal of Informetrics*, *11*(4), 1158–1174. https://doi.org/10.1016/J.JOI.2017.10.002

Leydesdorff, L., & Bornmann, L. (2020). "Interdisciplinarity" and "Synergy" in the Œuvre of Judit Bar-Ilan. *Scientometrics*, *123*(3), 1247–1260. https://doi.org/10.1007/s11192-020-03451-3

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics*, *13*(1), 255–269. https://doi.org/10.1016/j.joi.2018.12.006

Locatelli, B., Vallet, A., Tassin, J., Gautier, D., Chamaret, A., & Sist, P. (2021). Collective and individual interdisciplinarity in a sustainability research group: A social network analysis. *Sustainability Science*, *16*(1), 37–52. https://doi.org/10.1007/s11625-020-00860-4

Pinheiro, H., Vignola-Gagné, E., & Campbell, D. (2021). A large-scale validation of the relationship between cross-disciplinary research and its uptake in policy-related documents, using the novel Overton altmetrics database. *Accepted and forthcoming with Quantitative Science Studies*.

Rivest, M., Vignola-Gagné, E., & Archambault, E. (2021). Article-level classification of scientific publications: a comparison of deep learning, direct citation and bibliographic coupling. *Accepted and Forthcoming with PLoS ONE*. https://doi.org/10.1371/journal.pone.0251493

Rylance, R. (2015, September 16). Grant giving: Global funders to focus on interdisciplinarity. *Nature*, Vol. 525, pp. 313–315. https://doi.org/10.1038/525313a

Schneider, F., Buser, T., Keller, R., Tribaldos, T., & Rist, S. (2019). Research funding programmes aiming for societal transformations: Ten key stages. *Science and Public Policy*, *46*(3), 463–478. https://doi.org/10.1093/scipol/scy074

Science-Metrix. (2018). *Review of the Human Frontier Science Program 2018*. Retrieved from https://www.hfsp.org/node/12547#book/

Wang, J., Thijs, B., & Glänzel, W. (2015). Interdisciplinarity and Impact: Distinct Effects of Variety, Balance, and Disparity. *Plos One*, *10*, e0127298. https://doi.org/10.1371/journal.pone.0127298

Wang, Q., & Schneider, J. W. (2019). *Consistency and validity of interdisciplinarity measures*. https://doi.org/10.1162/qss_a_00011

Zuo, Z., & Zhao, K. (2018). The more multidisciplinary the better? - The prevalence and interdisciplinarity of research collaborations in multidisciplinary institutions. *Journal of Informetrics*, *12*(3), 736–756. https://doi.org/10.1016/j.joi.2018.06.006

A bibliometric strategy for identifying benchmark research units

Qi Wang[1] and Tobias Jeppsson[2]

[1] *qiwang@kth.se*
KTH-Royal Institute of Technology, KTH Library & Division of History of Science, Osquars backe 31, 100 44 Stockholm (Sweden)

[2] *tjep@kth.se*
KTH-Royal Institute of Technology, KTH Library, Osquars backe 31, 100 44 Stockholm (Sweden)

Abstract

While normalized bibliometric indicators are expected to resolve the subject-field differences between organizations in research evaluations, size still matters. Furthermore, research organizations, policymakers and research funding providers tend to use benchmark units as points of comparison for a given research center in order to understand and monitor its development and performance. In addition to monitoring and evaluations, the identification of comparable benchmark organizations can also be used to pinpoint potential collaboration partners or competitors. Therefore, methods to identify benchmark research units are of practical significance. However, few studies have investigated this problem. This study aims to propose a bibliometric method to identify benchmarks. We define a benchmark as a well-connected research environment, in which researchers work on similar topics and publish a similar number of publications compared to a given research center during the same period. Three essential attributes for the evaluation of benchmarks are research topics, output, and coherence. We apply this strategy to a Swedish research center, and examine the effectiveness of the method.

Introduction

Few studies in the field of research evaluation and research policy have focused on identifying benchmark research organizations. One reason for this is probably that normalized bibliometric indicators are used to deal with differences in research-field and publication output between research organizations, and this solution may be seen as sufficient. However, size still matters, both when evaluating the research performance of organizations that change in size over time, and since researchers at a larger organization are more likely to benefit from peer- and organizational effects. Field-normalized citation indicators that work well for large organizations and across coarse field-classifications may also be misleading for smaller organizations with a narrower research focus, where comparisons against relevant benchmarks may be more accurate.

A key aspect in identifying benchmarks is defining the relevant attributes for comparisons. Even so, we only found two studies on this topic. Carayol et al. (2012) propose a method for selecting peer universities and departments, and Andersen et al. (2017) develop a bibliometric-based approach for choosing proper benchmarks. Even though both these studies lack a concrete definition of benchmarks, their perceptions of a benchmark unit can be gleaned from their proposed approaches. The approach by Carayol and colleagues (2012) implies that the quantity of scientific production and its impact are the two essential attributes for a benchmark. In addition to these two attributes, Andersen and colleagues (2017) stress the importance of research topics in the selection of benchmarks. More specifically, they hold that “the topicality or subject profile” of a benchmark unit should be approximately similar to the treatment.

Based on the previous studies, benchmark units should consist of researchers who work on the same topics, publish a similar number of papers in relation to the unit under evaluation during the same period. In addition to these factors, we suggest that a benchmark also needs to form a well-connected and coherent research environment, to function as a useful benchmark unit. In this study, we exclude research impact suggested by Carayol et al (2012) from the primary identification of benchmark units, but it can still be useful as a soft attribute to select from potential benchmarks. Accordingly, the main attributes to consider when detecting benchmarks to research organizations should be (i) research topics, (ii) output, and (iii) coherence.

Methodology

In this section, we go through the three attributes for the identification of benchmarks in turn below.

Research topics

Delineating the research topics of an evaluative organization is one of the most important procedures for the identification of benchmarks. In this study, research topics were defined as publication clusters, using the clustering method of Waltman and Van Eck (2012, 2013) based on direct citation links. The reasons for choosing this strategy were summarized in Wang (2018). Our employed clustering system assigned around 36 million publications (articles and reviews) into 5,053 clusters. These publications were published from 1980 to 2019 and covered by the Web of Science (WoS) database. It has a similar scale compared to the one used in Leiden Ranking (2020), which consists of around 4,000 clusters[1].

The research topics of a treatment unit are, then, described by the clusters that include the unit's publications. After mapping the distributions of the 15 Swedish research centers, we conclude that the distributions of publications over clusters are usually highly skewed, as shown in Fig.1. Most of these centers are not interdisciplinary oriented. It is of little significance to delineate the research profile of a treatment group using the clusters that have an extremely small number of its publications. In addition to this, one should notice that the size of clusters varies greatly in our clustering system. During the 38-year period, the largest cluster consists of around 56,000 publications, whereas the smallest has only 500 based on the parameters we set. Therefore, for an evaluative group, it would be more reliable to consider its relative number of publications in each cluster to define its dominant research topics.

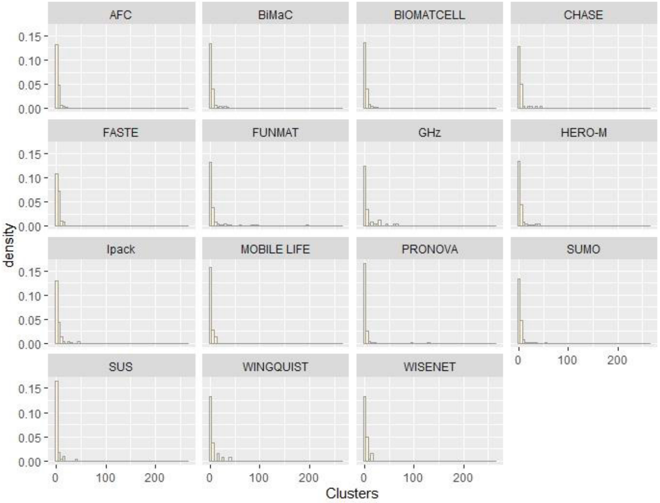


Figure 1. Distribution of publications over clusters for 15 research centers.

To be specific, let $p_{i,ts,\,te}^{k}$ denote the number of publications of evaluative unit i in cluster k from the year ts to te, then its total number of publications can be expressed as $P_{i,ts,\,te} =$

[1] More detailed information on the fields in Leiden Ranking is available at https://www.leidenranking.com/information/fields

$\sum p_{i,ts,te}^{k}$. Let $t_{ts,\,te}^{k}$ denote the total number of publications in cluster k from the year ts to te, and $a_{i,ts,\,te}^{k}$denote the share of publications for i in cluster k, then we have $a_{i,ts,\,te}^{k} = \frac{p_{i,ts,\,te}^{k}}{t_{ts,\,te}^{k}}$. To pinpoint dominant clusters for unit i, we intend to select the clusters with a high a. Let r denote the rank of a in decreasing order, and it hence satisfies $a_{i,ts,\,te}^{k(r-1)} \geq a_{i,ts,\,te}^{k(r)}$. Let $d_{i,ts,\,te}$denote the share of publications for unit i in the first nth clusters, then we have

$$d_{i,ts,\,te} = \frac{p_{i,ts,\,te}^{k(1)}+p_{i,ts,\,te}^{k(2)}+\cdots+p_{i,ts,\,te}^{k(n-1)}+p_{i,ts,\,te}^{k(n)}}{\sum p_{i,ts,te}^{k}} = \frac{\sum_n p_{i,ts,\,te}^{k(r)}}{\sum p_{i,ts,te}^{k}}. \tag{1}$$

To identify the dominant clusters, we require $d_{i,ts,\,te} \geq d_{min}$. In the meanwhile, we also need to avoid selecting the clusters with an extremely small number of unit i's publications, which requires $p_{j,ts,\,te}^{k(r)} \geq p_{min}$. By setting the parameter values, nth clusters can be selected to delineate the research profile for unit i.

Research output

After the selection of the dominant clusters, publications between year ts and te in the nth clusters can be aggregated into research organizations based on the affiliations of authors. Let $P_{j,ts,\,te}$ denote the publications of organization j in the nth clusters, which can be expressed as $P_{j,ts,\,te} = \sum_n p_{j,ts,\,te}^{k(r)}$. As discussed, for a research unit to be considered as a benchmark, it should have a similar number of publications with the treatment, that is $P_{j,ts,\,te} \in (P_{i,ts,\,te} - \Delta p,$ $P_{i,ts,\,te} + \Delta p)$.

Coherence

Researchers at a benchmark unit are expected to be connected to some extent. In this work, a weighted measure of the clustering coefficient (Opsahl & Panzarasa, 2009) has been applied to examine the extent to which researchers working at a benchmark unit are well connected. According to Opsahl and Panzarasa (2009), the clustering coefficient evaluates the total value of closed triplets in a weighted network in comparison with the total value of triplets. The higher the clustering coefficient, the more connected a network is. They have proposed several measures for calculating the weighted triplet value, including arithmetic mean, geometric mean, maximum and minimum, and further indicated the choice of the weighted measure should be based on the research question at hand. In our network, nodes represent researchers and links refer to the number of their collaborated publications. We assume that as long as researchers A and B, A and C have established collaboration relations respectively, B and C are then likely to collaborate. We hence choose the minimum measure. Let $h_{j,ts,\,te}$ denote the clustering coefficient of a benchmark unit, it should satisfy that $h_{j,ts,te} \geq h_{min}$.

In summary, a benchmark research unit should satisfy each of the following criteria: 1. $d_{i,ts,\,te} \geq d_{min}$ it should present similar research interests with the treatment; 2. $p_{j,ts,\,te}^{k(r)} \geq p_{min}$ it should have a certain amount of research output published in the main research topics of the treatment; 3. $h_{j,ts,te} \geq h_{min}$ it should be coherent.

Data

In this study, we applied the proposed method to identify benchmark units for some research centers that have been financed by Sweden's Innovation Agency (VINNOVA). Hero-m, at KTH-Royal Institute of Technology, was used as an example to further demonstrate our approach and to elaborate the results. This center aims "to develop tools and competence for fast, intelligent, sustainable and cost-efficient product development for Swedish industry.

Continuous scientific breakthroughs are exploited to enable design of materials from atomistic scales to finished products"[2]. Based on the information available on their webpage, we have collected 255 publications between 2007 and 2019, which belong to 62 clusters. As shown in Fig.1., the distribution of its publications over clusters is highly skewed and most of the publications are assigned in the top seven clusters. In addition to this, its coherence is 0.39.

Results

We used two sets of parameter values to examine the proposed method, as summarized in Table 1. The various combinations of parameter values would be useful to examine the sensitivity of the results. For the first set, we expect the first nth clusters, ranked by the share of Hero-m's publications over clusters, to have about 80% of Hero-m's total publications. Additionally, the dominant clusters are supposed to include no less than 5 Hero-m's publications. We further require the total publications of a benchmark in those dominant to be within the range between 0.7 and 1.3 times Hero-m's total publications. For the second set, Hero-m's publications in the dominant clusters should account for 60% of its total output and each dominant should include no less than 10 Hero-m's publications. Potential benchmarks are, then, the research organizations that have publications in these dominant clusters. More specifically, a benchmark should satisfy the criteria that its total publications are between 0.5 and 1.5 times Hero-m's publications, and its coherence is no less than 80% of Hero-m's values, which is 0.3.

Table 1. Two sets of parameter values

	Parameter set - 1	*Parameter set – 2*
d_{min}	212 (80%$P_{i,ts,\,te}$)	159 (60%$P_{i,ts,\,te}$)
p_{min}	5	10
Δp	53 (30%$P_{i,ts,\,te}$)	132.5 (50%$P_{i,ts,\,te}$)
c_{min}	0.39 ($c_{i,ts,\,te}$)	0.3 (80%$c_{i,ts,\,te}$)

The two sets of parameter values yielded 31 and 52 benchmarks respectively. For lack of space, Table 2 only presents the top ten most published benchmarks. The number of publications and coherence of each benchmark can be found in this table.

We acknowledge that our approach has a trade-off between precision and recall. Unfortunately, it is impossible to understand entire benchmarks for an evaluative unit, and hence the magnitude of the trade-off is very difficult to estimate. Nevertheless, we consider that precision is more important than recall in terms of the present research question. In other words, identifying appropriate comparable research units seems to be more meaningful in practical applications. It should be stressed that our approach is quite flexible, which allows researchers, policymakers, and other practitioners to adjust parameter values to detect benchmarks according to their specific purposes.

Table 2. Top benchmark research units for Hero-m

Identified benchmarks	*Parameter set*	*Output*	*Coherence*
Dept Nucl Engn & Radiol Sci, University of Michigan, US	set 1	341	0.47
Dept Mat Sci & Met, University of Cambridge, UK	set 1	330	0.39
State Key Lab Metastable Mat Sci & Technol, Yanshan University, PRC	set 1	330	0.42

[2] More detailed information regarding the Hero-m center is available at https://www.mse.kth.se/research/research-center/hero-m-2

Dept Mat Sci & Engn, Knoxville, University of Tennessee, US	set 1	299	0.46
Div Mat Sci, Argonne National Laboratory, US	set 1	291	0.59
Sch Mfg Sci & Engn, Sichuan University, PRC	set 1	280	0.43
Dept Mat Engn, Adv Mat Res Ctr, Islamic Azad University, IR	set 1	274	0.57
Inst Ind Sci, University of Tokyo, JP	set 1	266	0.69
Dept Mat Sci & Engn, Northern Illinois University (NIU), US	set 1	264	0.49
Dept Phys & Astron, Uppsala University, SE	set 1	256	0.53
Grad Inst Ferrous Technol, Pohang University of Science and Technology, KR	set 2	370	0.35
Tokai, Japan Atomic Energy Agency, JP	set 2	356	0.60
Los Alamos National Laboratory, US	set 2	343	0.38
Natl Die & Mold CAD Engn Res Ctr, Shanghai Jiao Tong University, PRC	set 2	332	0.53
Fac Phys & Appl Comp Sci, AGH University of Science and Technology, PL	set 2	320	0.44
Inst Adv Energy, Kyoto University, JP	set 2	308	0.53
National Insitute of Materials Science (NIMS), JP	set 2	302	0.57
Inst Appl Mat, Karlsruhe Institute of Technology, DE	set 2	289	0.55
State Key Lab Metastable Mat Sci & Technol, Yanshan University, PRC	set 2	284	0.44
China Iron & Steel Research Institute Group, PRC	set 2	269	0.40

Validation

In this section, we further examine the effectiveness of the proposed method.

Examining the consistency of research topics

The Latent Dirichlet Allocation (LDA) was employed to generate underlying research topics from the publications of each identified benchmark. In our work, the abstract of publications was used to generate topics. Some frequently occurred terms were considered as stop words. Since Hero-m is not a multidisciplinary research center with broad and diverse research topics, we determined to generate simply two topics for this center and each identified benchmark. Furthermore, for each topic, we listed the first 20 terms according to their probabilities of the inferred topics. The result of LDA for both treatment and benchmarks are available online[3]. Comparing the results in the two tables, the topics of Hero-m and its benchmarks are rather similar, focusing on properties, structure, and calculation of materials. But we have also noticed that some benchmarks have one of the topics regarding nuclear, ray, radiation, and irradiation, for instance, one of the research focuses at Pacific Northwest National Laboratory[4] is relevant to nuclear materials. On closer examination, we found Hero-m has indeed a small number of publications on this topic, such as Li and Korzhavyi's study in 2015 (*Interactions of point defects with stacking faults in oxygen-free phosphorous-containing copper*) and Xu and colleagues in 2020 (*Nuclear and magnetic small-angle neutron scattering in self-organizing*

[3] The results of LDA are available online https://kth.box.com/shared/static/1m4m66v8gnku4lzi4k1voobbl90gsqvy.docx

[4] More detailed information on Pacific Northwest National Laboratory are available at https://www.pnnl.gov/materials-science

nanostructured Fe1− xCrx alloys). For this reason, organizations with a focus on nuclear materials, such as Pacific Northwest National Laboratory, have been identified.

Examining the robustness of different coherence measures

In this study, we have chosen the minimum statistic as a measure of coherence. However, it is unclear if our results are robust to other coherence measures. We have therefore applied the four different measures, to calculate coherence for the benchmark units identified with the use of the second set of parameter values. The shows high consistency of various coherence measures, which Pearson correlation coefficients are all above 0.95.

Discussion and future work

Besides the validation tests mentioned in the previous section, we have also examined the performance of using the WoS subject categories as research topics, instead of publication clusters. However, 69% of Hero-M's publications belong to the category, Material Science Multidisciplinary, which is a rather broad research topic and cannot precisely describe the research focus of the center. Thus, we believe a finer classification system must be used for this type of study. However, using a classification that is too granular can instead run the risk of creating self-referential categories, meaning that most publications in individual clusters are from one individual research group. By analyzing the share of organizations in each cluster, we found that, only in 21 clusters, a single organization accounts for more than 40% of the total publications. In other words, the applied classification system does not seem to present a self-referential problem.

In the next step, we would like to examine the performance of this method to interdisciplinary research centers and also examine other methods to evaluate cluster profiles between the focal unit and potential benchmarks.

Acknowledgments

We would like to thank Lennart Stenberg for his helpful comments and suggestions. We are grateful to VINNOVA for their support of this study. We also thank reviewers for their comments on this paper.

References

Andersen, J. P., Didegah, F., & Schneider, J. W. (2017). *The necessity of comparing like with like in evaluative scientometrics: A first attempt to produce and test a generic approach to identifying relevant benchmark units.* In STI Conference Science and Technology Indicators Conference. Paris.

Hornik, K., & Grün, B. (2011). Topicmodels: An R package for fitting topic models. *Journal of statistical software*, 40(13), 1-30.

Katz, J. S. (2000). Scale-independent indicators and research evaluation. *Science and Public Policy*, 27(1), 23-36.

Li, Y., & Korzhavyi, P. A. (2015). Interactions of point defects with stacking faults in oxygen-free phosphorus-containing copper. *Journal of Nuclear Materials*, *462*, 160-164.

Opsahl, T., & Panzarasa, P. (2009). Clustering in weighted networks. Social networks, 31(2), 155-163.

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, 64(2), 372-379.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378-2392.

Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European physical journal B*, 86(11), 471.

Wang, Q. (2018). A bibliometric model for identifying emerging research topics. *Journal of the association for information science and technology*, 69(2), 290-304.

International mobility and gender balance in academia. The case of Norway.

Kaja Wendt[1] and Hebe Gunnes[2]

[1] *kaja.wendt@nifu.no*
NIFU Nordic Institute for studies in Innovation, Research and Education, Økernveien 9, 0653 Oslo (Norway)

[2] *hebe.gunnes@nifu.no*
NIFU Nordic Institute for studies in Innovation, Research and Education, Økernveien 9, 0653 Oslo (Norway)

Abstract

Female representation among students and graduates in higher education is growing internationally. This is a promising trend for achieving gender balance in top positions in academia. But there is still a long way to go, as women accounted for 24 per cent in top positions at European higher education institutions in 2016. In this article, we examine the influence of international recruitment on gender balance in Norwegian academia. We draw on data from the Norwegian Register of Research personnel, linked with population statistics from Statistics Norway. These data show that 32 per cent of the researchers at Norwegian higher education institutions in 2018 were born abroad. The share of foreign full professors has increased from 16 per cent in 2001 to 27 per cent in 2018, while for postdocs there has been an increase from 31 to 69 per cent. In terms of gender composition, a higher percentage of the foreign-born researchers are male. The incidence of international recruitment differs significantly across academic fields and positions. These dimensions are further analysed from a gender perspective in this paper.

Introduction

Many countries have put gender balance high on the research policy agenda to utilize the pool of talent and improving the quality of research. During the last decades the gender balance in research and higher education has improved, but imbalances persist and vary across countries, sectors, academic positions, and disciplines, as manifested in reports such as the *She Figures 2018* (EC 2019). Women are particularly underrepresented in top academic positions. Moreover, the lack of gender balance is especially strong in some of the STEM-fields (Science, technology, engineering, and mathematics). This has been a matter of concern by governments in many countries as it is believed that the quality of work and innovation in these fields may increase with more women. There are multifaceted arguments for increased gender balance, ranging from ethical, and economic to legal arguments. The European Commission highlights gender balance from the perspective of avoiding waste of talent (2010) and has recently again emphasised the importance of gender equality to strengthen European research and innovation (EC 2020). In the Horizon Europe Strategic plan 2021–2024 the integration of gender equality is highlighted as a cross-cutting issue, especially concerning global challenges in which gender differences play an important role that determine the societal relevance and quality of research and innovation outcomes.

The lack of gender balance has several explanations. In this paper we address the issue by investigating the role of international mobility and recruitment from a gender perspective. The internationalization of science has accelerated during the last decades, caused by factors such as increased access to physical and digital communication. In addition, the labor market for researchers has been increasingly internationalized with positions that are often open for applicants from abroad. Most vacant positions at Norwegian higher education institutions (HEIs) are advertised internationally.

International mobility of academics is supported by most Western countries, even though there is an asymmetry of mobility flows, some countries and regions are more attractive than others depending on historical, political, and economic factors. Higher demands for high-quality tertiary education worldwide are coupled with specific policies to promote student mobility within a geographic region (as is the case in Europe). Attracting mobile students and researchers is a way to tap into a global pool of talent and support the development of innovation and production systems across OECD countries. On average 10 per cent of first-time entrants into tertiary education were international students in 2018 (Education at a Glance 2020). Also in the EU mobility has been a cornerstone of the European Research Area (ERA) and its success is determined by highly skilled people who move across borders to where their talent can be best employed; ERA Priority 3 (EC 2020).

Still there is little knowledge about the consequences of increased international mobility of researchers on gender balance. The EU MORE 3-survey (EC, 2017) among 10,000 European researchers revealed that women are more mobile at early stages of their career, while men are more mobile later. Other studies have looked into reasons for low representation of women at higher career-levels in academia by investigating driving forces for geographic mobility which influence men and women differently (Leemann 2010, Canibano 2016).

We have found no other studies that look at the impact of international mobility of researchers on gender balance within countries. Does such mobility improve the domestic gender balance or not? Obviously, this issue will arise differently among countries. And in a sense, it will be a zerosum game: Influx of female researchers from one country to another reduces the population of women in first country and increases it in the second, recipient country.

In our study we look at incoming mobility[1] to Norway. Norway is a small country which contributes to 0.4 per cent of global research and development (R&D) (OECD 2020a). Attracting the best talents and promoting diversity have been high on the political agenda when international recruitment is discussed in Norway (Vabø 2020). Norway may be an interesting country for analyzing this issue because the country is a small and open country with strong research links to the rest of Europe and the USA and facilitates the mobility and work of foreign researchers. As part of the EEA agreement, all EU citizens are allowed to work in Norway and also researchers from most other countries can easily get work permissions.

Approximately two thirds of the full professors in Norway obtain their professorship through a "professor by competence" system[2]. This implies that the recruitment of associate professors, as the recruitment pool for full professorships, is crucial to obtain gender balance in the top positions in Norwegian HEIs. In the current academic career system, the postdoc position is an important step on the career ladder to become an associate professor. We have thus selected these three positions for further analysis. We will also examine the gender balance among researchers born in Norway or abroad.

The following main research questions are addressed: a) What is the gender balance among the native and foreign researches across fields and over time? b) Has the international recruitment hampered the process of achieving gender balance in the Norwegian HE-system? In the study we will look at incoming mobility to Norway for the crucial career steps of postdoc, associate professor and full professor in different fields of research. The analysis covers 2007–2018.

[1] Our dataset contains limited information about outgoing mobility of Norwegian-born researchers, as we do not have information about Norwegians studying for a PhD abroad.

[2] The Norwegian scheme of promotion to professor after a competency assessment was introduced in 1993 and is part of the Act relating to Universities and University Colleges. The purpose was among other things to create a more fair system. Previously you could only become a full professor if you applied for, and got, a vacant professorship. A better career system, improved quality of higher education and research, and to increase the number of female professors were other goals. Today 66 per cent of new professors have their promotion through this system.

Data and method

The study is based on the Norwegian Register of Research personnel[3], operated by NIFU, combined with data of population statistics from Statistics Norway[4]. The dataset includes variables on country of birth, gender, position, and field of research. In the analyses each individual is classified as either Norwegian (native) or foreign based on country of birth. We provide descriptive overviews of changes in the career systems over time and by field according to these dimensions in a gender perspective. As a first step, we present an overview of the current situation in Norway (2018). Then we analyse individuals who have achieved new positions as full professors and associate professors. This population of researchers are of particular interest as they reflect the most recent developments.

Results

In Norway, international recruitment is highest for the temporary positions, such as postdoc and research fellow. In 2018, a total of 33 per cent of the researchers at Norwegian HEIs were born outside Norway. Approx. 70 per cent of full professors and associated professors were born in Norway, while at postdoctoral level the picture is opposite; more than 70 per cent of the postdocs were born abroad (Table 1). Almost half of the researchers at Norwegian HEIs were women (49 per cent). Among researchers born abroad the share was 45 per cent, while the share among native Norwegian researchers was 52 per cent. Women accounted for 31 per cent of the full professors and 49 per cent of the associate professors, while 45 per cent of postdocs and 54 per cent of the research fellows were female. The percentage of women is almost identical for native and foreign full professors (one percentage point difference), while there is a larger difference for postdocs and temporary research positions (11 percentage points).

Table 1. Full professors, associate professors and postdoctors by immigrant status[1] and share of women in Norwegian HEIs. 2018 (percentages).

Position	*Status*	*Share of total*	*Share of women*	*Share of men*	N=
Full professor	Native	70 %	31 %	*69 %*	3 021
	Foreign	30 %	30 %	*70 %*	1 096
Associate professor	Native	69 %	51 %	*49 %*	3 151
	Foreign	31 %	45 %	*55 %*	1 305
Postdoctor	Native	28 %	53 %	*47 %*	444
	Foreign	72 %	42 %	*58 %*	1 130

[1]Immigrant status relates to country of birth.

Source: NIFU, Statistics Norway

Our data show that in Norwegian HEIs the share of foreign researchers varies clearly between field of research. In 2018, 30 per cent of the researchers in humanities and the arts were born abroad, while social sciences had 23 per cent foreign researchers. The highest share of foreign researchers was found in natural sciences and engineering and technology (hence 49 and 50 per cent). In medical and health sciences, 27 per cent of the researchers were born abroad.

A closer look at full professors in 2018 shows that in humanities and the arts 36 per cent were female (38 per cent of the natives and 34 per cent of those born abroad), while in social sciences, 33 per cent of the native and 37 per cent of the foreign full professors were women. In

[3] The Register of Research personnel is part of the official Norwegian R&D statistics. The register covers researchers/university graduated personnel that participated in R&D at Norwegian HEIs, as well as research institutes and health trusts. The register includes information on position, age, gender, educational background and workplace. Personnel data is retrieved from the administration of the R&D-performing institutions per October 1st, and the registry goes back to the 1960s. Mobility, gender balance and career paths of the academic staff in Norway has thus been monitored through several decades through this rather unique database.

[4] Information on country of birth is retrieved from Statistics Norway. We also have information about which country foreign-born researchers immigrated from, if other than birth country.

engineering and technology, only 14 per cent of the full professors were women. The share of women was higher for the foreign full professors (16 per cent), compared to 13 per cent for the natives. We see the same trend for natural sciences, with 21 per cent female full professors born abroad, and 19 per cent of the native.

Table 2. Gender balance for full professors and associate professors by immigrant status[1], at Norwegian HEIs. 2007–2018[2].

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Full professor (N)	2 964	3 029	3 089	3 186	3 285	3 397	3 559	3 673	3 766	3 884	4 042	4 119
Female full professors (%)	18 %	19 %	20 %	21 %	23 %	24 %	25 %	26 %	27 %	28 %	30 %	31 %
Associate professor (N)	2 936	3 010	3 086	3 164	3 157	3 250	3 356	3 545	3 744	3 976	4 239	4 456
Female associate professors (%)	35 %	36 %	38 %	38 %	39 %	40 %	42 %	44 %	46 %	47 %	48 %	49 %

[1]Immigrant status relates to country of birth.
[2]Note that a number of new HEIs were included in the Register of Research Personnel in 2007, 2013 and 2017. Some of the new professors in these years are new entries in the database but held professor positions previously.
Source: NIFU, Statistics Norway

The female representation among the tenured personnel has grown steadily over time, as shown in Table 2, with figures for full professors and associate professors in Norway. The percentage of women has increased from 18 per cent for full professors in 2007 to 31 per cent in 2018. For associate professors, there has been an increase from 35 to 49 per cent in the same period.

In order to analyze this question further, we have looked specifically at researchers in new positions. This shows that the extent of international recruitment has increased significantly over time and the gender balance has improved. In 2007, 19 per cent of the new full professors were female and native, whereas 9 per cent were female and foreign. 11 years later, the share of female new full professors had increased to 42 per cent; 28 per cent native women and 14 per cent foreign women. For associate professors, there has been a notable increase in the share of foreign women (11 to 19 %) and men (14 to 21 %), while there has been a decrease in the share of native med (39 to 23 %). The share of native women has been relatively stable (34 to 37 %). The data shows that Norwegian men are still the largest group of new full professors, but the share of women, both native and foreign, is growing steadily. Native men are outnumbered by native women in the associate professor position, and they might soon be outnumbered by foreign-born men. A closer look at the recruitment of new full professors, see figure 1, shows that the share of foreign full professors has increased in all fields except medicine and health sciences. There is a significantly higher share of new male foreign full professors over foreign female, again with the exception of medicine and health sciences. At the same time, the share of new male native full professors seems to be decreasing in all fields, while the share of native female full professors is relatively stable in the STEM-fields and increase in medicine and health sciences. However, the share of men among the new full professors is still higher than the share of women in most fields, and especially in the STEM-fields.

When analyzing the population of native and foreign new full professors separately, the overall female proportion is 31 and 30 per cent, respectively in 2015–2018 (Table 3). However, at field levels, there is a better gender balance among the foreign full professors than the native in several of the fields, although the difference is not very large.

There is clearly a higher share of women, both native and foreign, among the new associate professors than it was for the full professors. This applies to all fields, but the differences were highest in 2015–2018 for engineering and technology and medicine and health sciences. For the first field, it was especially the share of foreign female associate professors that was higher, for the latter it was the share of native female associate professors. These results are not shown, but will be examined closer in the full paper.

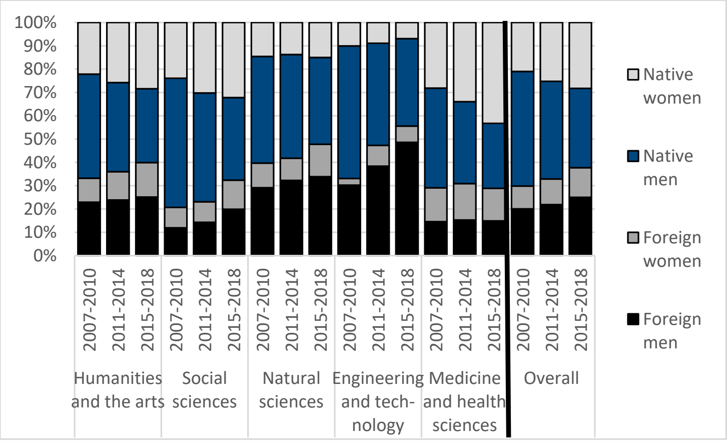


Figure 1 New full professors at Norwegian HEIs by field of R&D[1], gender and immigrant status[2]. 2007–2018 (percentages).

[1]Natural sciences include veterinary and agricultural sciences.
[2]Immigrant status relates to country of birth.
Source: NIFU, Statistics Norway

Table 3. Native and foreign[1] women as a share of all new full professors and native/foreign professors bye field of science[2]. 2007–2018. Per cent.

			2007-2010[3]	2011-2014[3]	2015-2018[3]
Overall	Native women	Proportion of all new full professors	21 %	25 %	28 %
		Proportion of native full professors	20 %	25 %	31 %
	Foreign women	Proportion of all new full professors	10 %	11 %	13 %
		Proportion of foreign full professors	26 %	29 %	30 %
Humanities and art	Native women	Proportion of all new full professors	22 %	26 %	28 %
		Proportion of native full professors	27 %	31 %	38 %
	Foreign women	Proportion of all new full professors	10 %	12 %	15 %
		Proportion of foreign full professors	30 %	32 %	34 %
Social sciences	Native women	Proportion of all new full professors	24 %	30 %	32 %
		Proportion of native full professors	21 %	27 %	33 %
	Foreign women	Proportion of all new full professors	9 %	9 %	12 %
		Proportion of foreign full professors	32 %	36 %	37 %
Natural sciences	Native women	Proportion of all new full professors	15 %	14 %	15 %
		Proportion of native full professors	14 %	16 %	19 %
	Foreign women	Proportion of all new full professors	11 %	10 %	14 %
		Proportion of foreign full professors	17 %	19 %	21 %
Enginee-ring and technology	Native women	Proportion of all new full professors	10 %	9 %	7 %
		Proportion of native full professors	9 %	11 %	13 %
	Foreign women	Proportion of all new full professors	3 %	9 %	7 %
		Proportion of foreign full professors	15 %	15 %	16 %
Medicine and health sciences	Native women	Proportion of all new full professors	28 %	34 %	43 %
		Proportion of native full professors	25 %	35 %	44 %
	Foreign women	Proportion of all new full professors	15 %	16 %	14 %
		Proportion of foreign full professors	38 %	45 %	44 %

[1] Foreign relates to country of birth.
[2] Natural sciences include veterinary and agricultural sciences.
[3] Share of native/foreign full professors is calculated for the last year of the period.
Source: NIFU, Statistics Norway

Discussion

This study has shown that the Norwegian HE-system has undergone rapid changes in terms of internationalization, where the share of foreign researchers has grown considerably over the last years, especially at lower levels of the academic career ladder and especially within the STEM-fields. However, among foreign researchers the share of women is lower (45 %) than among the native Norwegians (52 %) and this might suggest that internationalization has slowed down the development towards gender balance.

However, there are significant differences in gender balance by field of research. Moreover, international recruitment is particularly prevalent in engineering and technology, where the gender balance is most skewed. Thus, the overall figures are affected by these patterns. When the issue is analysed at the level of fields, a different picture emerges. The female proportion of new foreign professors (2015-2018) is higher than the female proportion of native professors in all fields, except humanities and arts, while it is equal in medicine and health. Contrary to what would be expected from the overall results, this shows that the international recruitment has in fact contributed positively to the gender balance in Norway in the majority of the fields. The final analysis will include a more detailed mapping of both associate professors and postdocs. We will also look into statistics on applicants for vacancies at the Norwegian HEIs.

Acknowledgments

We would like to thank our colleague, Dag W. Aksnes who is the project leader of the BALANSEkunst project at NIFU, for comments on this early draft of the paper.

References

Canibano, C., Fox. F M. & Otamendi, F. J (2016): Gender and patterns of temporary mobility among researchers. *Science and Public Policy*, 43(3), 320–331, doi: 10.1093/scipol/scv042

European Commission (2019): *She Figures 2018.* Luxembourg: Publications Office of the European Union, 2019

European Commission (2020): A new ERA for Research and Innovation. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS: Brussels 30.09.2020, COM(2020) 628 final

European Commission (2021): Horizon Europe Strategic Plan (2021–2024), Directorate-General for Research and Innovation

Leeman, R. J. (2010). Gender inequalities in transnational academic mobility and the ideal type of academic entrepreneur, Discourse: Studies in the Cultural Politics of Education Vol. 31(5), December 2010, 609–625

OECD (2020). *Education at a Glance 2020*, OECD

OECD (2020a): Main Science and Technology Indicators (MSTI) 2020/1.

OECD (2021), OECD Science, Technology and Innovation Outlook 2021: Times of Crisis and Opportunity, OECD Publishing, Paris, https://doi.org/10.1787/75f79015-en

Vabø, A. (2020). The relevance of international recruitment for working conditions in higher education and research, Norwegian Journal of Sociology, Årgang 4, nr. 1-2020, s. 34-41, ISSN online: 2535-2512, DOI: https://doi.org/10.18261/issn.2535-2512-2020-01-03 Årgang 4, nr. 1-2020, s. 34–41 I

Hierarchical topic tree: A hybrid model comprising network analysis and density peak search

Mengjia Wu[1] and Yi Zhang[2]

[1] *Mengjia.Wu@student.uts.edu.au*
Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney (Australia)

[2] *Yi.Zhang@uts.edu.au*
Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney (Australia)

Abstract

Topic hierarchies can help researchers to develop a quick and concise understanding of the main themes and concepts in a field of interest. This is especially useful for newcomers to a field or those with a passing need for basic knowledge of a research landscape. Yet, despite a plethora of studies into hierarchical topic identification, there still lacks a model that is comprehensive enough or adaptive enough to extract the topics from a corpus, deal with the concepts shared by multiple topics, arrange the topics in a hierarchy, and give each topic an appropriate name. Hence, this paper presents a one-stop framework for generating fully-conceptualized hierarchical topic trees. First, we generate a co-occurrence network based on key terms extracted from a corpus of documents. Then a density peak search algorithm is developed and applied to identify the core topic terms, which are subsequently used as topic labels. An overlapping community allocation algorithm follows to detect topics and possible overlaps between them. Lastly, the density peak search and overlapping community allocation algorithms run recursively to structure the topics into a hierarchical tree. The feasibility, reliability, and extensibility of the proposed framework are demonstrated through a case study on the field of computer science.

Introduction

The last decades have witnessed a great accumulation of scientific documents, resulting in information overload for researchers. Aiming to improve this situation, a substantial number of bibliometric studies on topic extraction, knowledge mining, and text analytics have been undertaken, each looking for efficient ways to extract information from textual data and concise ways of presenting the knowledge found (Ba et al., 2019; Qian et al., 2020; Song et al., 2016; Wu et al., 2020; Zhang et al., 2018; Zhang et al., 2017). What many of those studies have shown is that, one, organizing research topics into curated hierarchical structures is an excellent way of quickly conveying a great deal of knowledge about the composition of a research field to those who are unfamiliar with it, and, two, constructing these arrangements is nontrivial and highly challenging. While very broad overviews of a field are not particularly difficult to generate, creating interactive topic maps that show fields at different and especially fine levels of granularity and disentangling the rising complexities of inter-/multi- disciplinary studies is another story altogether. In fact, all but the most rudimentary techniques still rely heavily on expert knowledge.

That said, advancements in natural language processing (NLP) are reducing this dependence, with methods capable of automatically identifying and stratifying the thematic concepts found in a dataset of literature. Among these methods, hierarchical latent Dirichlet allocation (hLDA) (Blei et al., 2010) is especially well-known. However, there are a couple of aspects of hLDA that could be improved. These include sometimes weak associations between the generated parent and child topics; and internal unigram incoherence within topics (Qian et al., 2020; Xu et al., 2018); a propensity to represent each topic as a conglomeration of unigrams and probabilities; and a tendency to label topics with appropriate names, which reduces the interpretability of the results. There are also alternative approaches of building topic hierarchies, such as taxonomy identification (Shang et al., 2020), ontology construction (Wong et al., 2012), and knowledge graphs (Yang et al., 2017). But, despite substantial efforts to the contrary, these techniques inevitably suffer from either an excessive number of parameters that need to be fine-tuned and/or issues with creating clean partitions between topics. Hard clustering algorithms like K-means or non-negative matrix factorization (Qian et al., 2020; Zhang et al., 2018), which most of these techniques are based on, struggle to find clear divisions between topics with high levels of overlap, convergence, or interactivity – characteristics that typify the process of scientific development.

Aiming to solve these issues, we propose a novel framework called *Hierarchical Topic Tree* (HTT) that operates in a recursive manner to reveal the topic hierarchies within a set of documents. The framework comprises a term co-occurrence network and two algorithms: DPS, a density peak search algorithm, modified to work with networks; and OCA, an overlapping community allocation algorithm. We assume that every topic consists of a core term, which becomes the topic's label, and a set of affiliated terms. Applying the density peak search algorithm to a term co-occurrence network reveals the density peak terms that meet some specific criteria for being used as a topic's label. The terms associated with every core topic term, i.e., the affiliated terms, are then determined and partitioned by the overlapping community allocation algorithm, which means terms can be assigned to multiple topics. These two steps are run recursively on partitioned subnetworks to identify deeper hierarchies in the term co-occurrence network until no core topic terms (topic labels) are found. To demonstrate the practical workings of the HTT framework, we conducted a case study on 6,267 academic articles published in the expansive field of computer science. The final results show a tree with six main branches and 120 sub-branches in a complex, but cohesive, hierarchical structure. The three main contributions our work makes include: 1) a density peak search algorithm that identifies and labels the topics in a corpus; 2) a community allocation algorithm that recognizes topic overlaps, which may indicate knowledge convergence; and 3) a model that requires two hyperparameters – a density threshold and an overlap threshold, which makes the process of tuning parameters easy and the model adaptable to a variety of cases.

The rest of this paper is organized as follows. The Related Works section next gives a brief review of the work on topic hierarchy identification. The Methodology section sets out the details of our proposed methodology. The Case Study section follows, presenting the data, results, and empirical insights derived from the computer science domain case. We then wrap our study with a conclusion, the study's limitations, and future directions of research.

Related Works

Hierarchies are instinctive, basal structures to humans that naturally aid our sensemaking of scientific knowledge composition. Blei pioneers the automation of topic hierarchy identification by developing the two perhaps most renowned algorithms in identifying topic hierarchies – the Chinese restaurant process (CRP) (Blei et al., 2004) and hierarchical latent Dirichlet allocation (hLDA) (Blei et al., 2010). However, the efficacy of the hLDA model largely depends on the pre-processing quality and may generate unsatisfactory results otherwise (Qian et al., 2020; Xu et al., 2018). The latter works pay efforts to modify topic hierarchy identification from different perspectives, including introducing the idea of recursive hierarchy detection (Wang et al., 2013), involving distance-dependent discrepancies for the CRP (Song et al., 2016), adding external ancillary information (Shang et al., 2020; Wang et al., 2015; Xu et al., 2018), and using alternative topic partition method like non-negative matrix factorization (Qian et al., 2020). But those studies either suffer from the need for a pre-defined tree structure or the lack of a labeling strategy. In practical terms, hierarchical structures vary hugely from discipline to discipline, especially for disciplines of vastly different forms, such as biomedicine versus artificial intelligence. As for the topic labeling strategy, most bibliometric approaches constitute topics as a set of semantically similar terms or records (Colavizza & Franceschet, 2016; Hou et al., 2018; Porter et al., 2020). Similarly, in mainstream topic modeling approaches, a topic is not represented with one all-encompassing label but, rather, as a bag of words or phrases and their corresponding probabilities. With both approaches, one still has to manually dive into the specific words, phrases, or even documents to infer the broad subject matter of the topic and decide on a name.

Density peak clustering was first proposed in *Science* by Rodriguez and Laio (2014). It is based on the premise is that the center of a cluster is more densely packed than the surrounding regions and that areas of high density tend to be relatively far apart. As a clustering method, density peak search has proven to be very fast and quite accurate. Compared to traditional K-means or density-based clustering algorithms like DBSCAN, density peak searches identify cluster centroids purely based on the one characteristic of density. There are no additional parameters and no multiple iterations, which means the clustering process is extremely efficient and highly robust to parameter selection. Du et al. (2016) have since improved this method by using average K-nearest neighbor (KNN) density to emphasize the importance of local density instead of the circle radius approach used originally. This notion of density accords with the characteristics a topic label should have in that a highly representative topic label will be strongly connected to its affiliated terms but as different as possible from other topic labels. This parallel motivated our idea to automatically name topics through a KNN-modified density peak-based clustering algorithm.

Methods

Concept definitions and problem formulation

Definitions of the main concepts referred to in the methodology are as follows.

- Topic term: Nominal phrases extracted from scientific documents via a series of natural language processing and cleaning steps.
- Topic: A set of topic terms with their corresponding probabilities headed by a core topic term. Term overlaps under the same parent topic are allowed for different topics.
- Hierarchical topic tree (HTT): HTT is both the name of our methodology framework and the final output. As an output, an HTT is a tree-structure that consists of topic nodes residing on different layers of a tree, as illustrated in Figure 1. The length from the root node to the nodes on the deepest layer is called the tree depth. A higher layer topic is a parent topic, and its connected topics in lower layers are called child topics. Child topics under the same parent topic are siblings. The associations between a parent and child topic are assumed to be stronger than the associations between siblings.

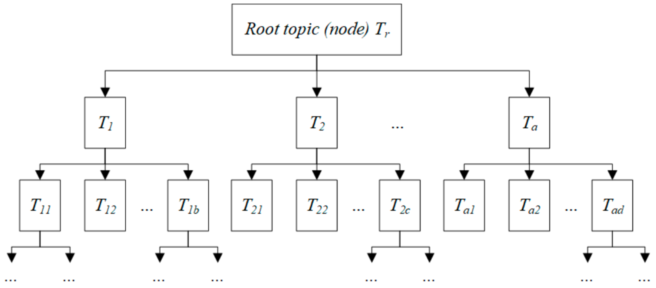


Figure 1. An example HTT

Problem formulation: The study aims to: 1) identify research topics with different granularities and construct a topic tree automatically from a collection of scientific documents; 2) label every topic with an appropriate name; and 3) detect topic overlaps.
HTT accomplishes these goals through the following steps.

Term clumping and network construction

The process begins by extracting topic terms from a corpus of documents. This is done with VantagePoint[1] and a term clumping process (Zhang et al., 2014). With the extracted terms in hand, the next step is to construct a weighted co-occurrence network of topic terms, denoted as $G = (V,\ E)$. V is the set of nodes representing the extracted topic terms, and E is the set of edges representing term co-occurrence. The graph is formulated according to the following equation:

$$w_{V_iV_j(i\neq j)} = \begin{cases} \dfrac{1}{CF(V_i, V_j)} & if\ V_i\ and\ V_j\ co-occur\ in\ at\ least\ one\ document \\ 0 & otherwise \end{cases}$$

where $w_{V_iV_j(i\neq j)}$ is the edge weight of $E_{V_iV_j(i\neq j)}$ and $CF(V_i, V_j)$ is the co-occurrence frequency of V_i and V_j.

Density peak search (DPS)

This algorithm is designed to identify core terms for topic labels. The primary concern when applying density peak clustering to network data is finding appropriate proxies for the distance and density measurements. Bai et al. (2017) use r-step topological distance as a proxy. However, this strategy necessitates a redundancy parameter r and a weighted parameter t, both of which need to be fine-tuned and both of which reduce the model's adaptability. Therefore, we opted to develop a new distance proxy, although still based on the topological distance between nodes:

$$d_{V_iV_j} = \begin{cases} w_{V_iV_j} & if\ V_i\ and\ V_j\ are\ connected \\ SPL_{V_iV_j} & if\ \ V_i\ and\ V_j\ are\ unconnected\ a\ path\ exits\ between\ them \\ \text{NA} & if\ no\ path\ exists\ between\ V_i\ and\ V_j \end{cases}$$

where $SPL_{V_iV_j}$ is the length of the shortest path from node V_i to V_j.

Generally, the co-occurrence network of high-frequency terms is fully connected, which means there will be at least one path from V_i to V_j. Hence, using the proposed new distance proxy, the kernel local KNN density and distance to the nearest denser point of every term can be calculated as:

$$\rho_{V_i} = \exp\left(-\frac{1}{K}\sum_{j\in KNN(V_j)} d(V_i, V_j)^2\right)$$

[1] More details could be found at www.vantagepoint.com.

$$\delta_{V_i} = \begin{cases} \max\limits_{V_j}(d_{V_iV_j}) & if\ \rho_{V_i} = \max(\rho_{V_i}) \\ \min\limits_{V_j\,\epsilon\,V_{\rho_{V_j}>\rho_{V_i}}}(d_{V_iV_j}) & otherwise \end{cases}$$

In the few cases where the co-occurrence network includes several unconnected components, we will generate a virtual root node for the final HTT. Then each component will be processed separately as a branch of the virtual root node.
The original DPC algorithm identifies the cluster centroids with higher values of ρ and δ by observing the $\rho - \delta$ plot. However, when applying this algorithm to a real-world dataset, the boundaries of centroids and other terms are not always that clear. Therefore, in HTT, these selection criteria are quantitative. V_c denotes the potential centroids of all the communities, and the criteria for selecting the final centroids are formulated as follows:

1) Density peak: The selected centroids should be density peaks, denoted as:

$$\rho_{V_c} = \max_{V_i \in KNN(V_c)} \rho_{V_i}$$

in which $KNN(V_c)$ denotes the K-nearnest neighbor nodes of V_c.

2) Sparsity: To guarantee the identified centroids are sparse to each other, we set the node's distance to its parent node as a quantitative minimum threshold, which also indicates the associations of child nodes are weaker than the associations with their common parent node. This criterion is expressed as follows:

$$\delta_{V_c} > d_{V_rV_c}$$

in which V_r denotes the parent node of V_c.

Initially, there is no root node to measure whether a node meets Criterion 2). Hence, we will only use Criterion 1) to identify root nodes. If only one node meets criterion 1), it will automatically become the root node. Otherwise, a virtual root node is generated, and the n identified nodes would become children to the virtual root.

Overlapping community allocation (OCA)
The next step is to distinguish overlapping topics between communities and ensure they are given proper multiple assignments. Thus, every node is assigned a probability vector $p_{V_i} = \{p_{i,1,} p_{i,2,} p_{i,3,} \dots, p_{i,n}\}$, which reflects the probabilities that V_i belongs to core terms identified. Specifically, the probability that node V_i belongs to a community (topic) with the core term V_c is calculated as follows:

$$p_{i,c} = \frac{\min\limits_{V_j \in V} d(V_j, V_c)}{d(V_i, V_c)}$$

In disjoint community allocation, node V_i will be exclusively allocated to its closest centroid c if $c = argmax_t\{p_{i,t}, t = 1,2,3, \dots, n\}$. However, our aim is to allocate a node to more than one potential community with high probabilities. Hence, we employ an overlap threshold σ to decide multiple communities the node V_i could belong to. The rule applied is that if $\frac{p_{i,t}}{p_{i,c}} > \sigma$, node V_i will be assigned to both community t and c. The output of this step is n overlapping communities with their assigned terms and probabilities.

Recursive hierarchy detection

The previous steps partition the network into n subnetworks, with each subnetwork comprising a core topic term and representing a sibling topic on the second layer. To extend the hierarchy into deeper layers, new subcommunities are detected by recursively applying the modified DPS and OCA algorithms to the partitioned subnetworks. When partitioning the parent networks into subnetworks, terms that belong to more than one topic, i.e., community overlaps, are excluded. This is because our approach aims at revealing hierarchies that exclusively belong to the parent topic. The recursive loop ends when no further core topic terms are detected in any subnetwork or the term number in the subnetwork is less than K.

The output of this step is the final HTT, with each node represented by a core topic term and linked to a set of terms. Topic overlaps containing terms shared by sibling topics are detected as well. This recursive process is illustrated in Figure 2, where each color represents a different stratum in the hierarchy. From top to bottom, the HTT has a root topic and one or multiple layers of topics generated by the iterations of DPS and OSA algorithms. Topics generated in the same iteration are siblings to each other and share a mutual parent topic.

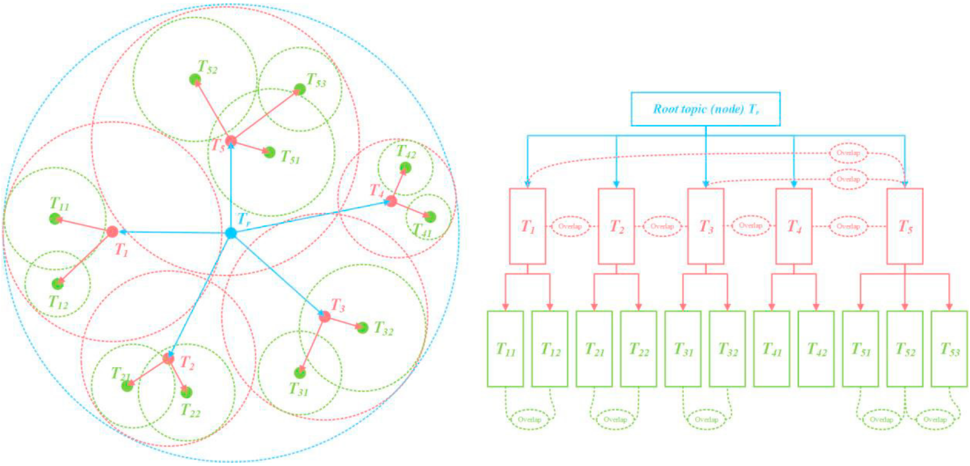


Figure 2. The recursive process of hierarchy construction

Methodology evaluation

According to criteria from previous studies, a well-curated hieratical topic structure should meet at least two characteristics: semantically coherent topics and high-quality parent-child topic relationships (Qian et al., 2020; Shang et al., 2020; Xu et al., 2018). Hence, we designed two indicators - topic coherence and parent-child association index (PCAI) to quantify the two characteristics. Additionally, we calculate the weight loss ratio of network edges to measure the information loss in the HTT process. Please note that the topics mentioned in this section contain overlapping terms, the association strength between two topics means the total sum of the edge weight's reciprocal of the pairwise terms from the two topics, the internal topic association of a topic strength refers to the total sum of the edge weight's reciprocal of pairwise terms from the topic itself.

- Topic coherence: Previous studies employ pointwise mutual information (PWI) to measure the topic coherence, but we consider it does not provide an intuitive and universal measure of topic coherence because its value range is -∞ to +∞ and its values vary hugely in multiple studies (Qian et al., 2020; Wang et al., 2013; Xu et al., 2018). Hence, in the current study, we measure the coherence of a topic via calculating the proportion of its total internal association strength against its total association strength with itself and its siblings.

$$Coherence_{T_i} = \frac{1}{|T_i|} \sum_{V_m \in T_i} \frac{\sum_{V_n \in T_i, m \neq n} CF(V_m V_n)}{\sum_{T_j \in children(parent(T_i))} \sum_{V_k \in T_j} CF(V_m V_k)}$$

- Parent-child association index (PCAI): This indicator is only applied to parent nodes in the final HTT (including the virtual root node if it exists). For every parent node, the PCAI equals the ratio of the total pairwise association strength among its children topics over the total association strength of itself and all children topics subtracted by 1.

$$PCAI_{T_i} = 1 - \frac{\sum_{T_m, T_n \in children(T_i), m \neq n} \sum_{V_p \in T_m^K, V_q \in T_n^K} CF(V_p, V_q)}{\sum_{T_j \in children(T_i)} \sum_{V_x \in T_j^K, V_y \in T_i^K} CF(V_x, V_y)}$$

- Information loss index: This index measures the overall information loss when the co-occurrence network being transformed into a hierarchical tree structure. The smaller value of information loss reflects the model's better performance of retaining information.

$$Information\ loss\ index_{T_i} = \frac{\sum_{T_m, T_n \in children(T_i), m \neq n} \sum_{V_p \in T_m, V_q \in T_n} CF(V_p, V_q)}{\sum_{V_x \in T_i, V_y \in T_i, x \neq y} CF(V_x, V_y)}$$

Case Study: The hierarchy of research topics in computer science

To demonstrate the methodology, we conducted a case study on the field of computer science, decomposing its many and varied research interests into topic hierarchies.

The corpus comprised 6,267 highly-cited papers published between 2010 and 2021 retrieved from the Web of Science (WoS) Core Collection database spanning the mainstream research topics regarding this domain. WoS is a well-curated multidisciplinary database with 74.8 million scientific publications from over 21,100 journals. Category information is assigned to every journal, and articles with the top 1% of citations received per field are flagged[2]. The search strategy used to assemble the corpus was as follows:

(WC = "Computer Science") AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article), Refined by: ESI Top Papers: (Highly Cited in Field), IC Timespan=2010-2021
WC: Web of Science Category;

Data pre-processing

Before applying our methods to the dataset, we ran VantagePoint's natural language processing (NLP) function to extract the raw words and phrases from the titles and abstracts. We then executed a term clumping process that removes noise and consolidates synonyms to arrive at a final list of topic terms. From this list, we selected terms with a frequency greater than 2. The

[2] https://clarivate.com/webofsciencegroup/solutions/essential-science-indicators/

stepwise cleaning results are given in Table 1. The final output was a term co-occurrence network consisting of 2,134 terms.

Table 1. Stepwise cleaning results

Step	*Description*	*# Terms*
1	Raw terms retrieved with NLP	132,846
2	Consolidated terms with the same stem, e.g., “information system” and “information systems”	116,898
3	Removed spelling variations, removed terms starting/ending with non-alphabetic characters, e.g., “Step 1” or “1.5 m/s”, removed meaningless terms, e.g., pronouns, prepositions, and conjunctions	114,459
4	Removed general single-word terms, e.g., “information” *	96,245
5	Consolidated synonyms based on expert knowledge, e.g., “co-word analysis” and “word co-occurrence analysis”	84,828
6	Eliminated all terms occurring less than 5 times	2,134

*Note: Given that most single-word terms take on additional context when used in multi-word phrases, e.g., “information” vs. “information systems”, we opted to remove generic single-word terms. Further, some multi-word terms were consolidated into a single-word form in Step 2 (e.g., “classification method” became “classification”). Non-general single-word terms were retained.

Parameter selection

Before generating the HTT, we selected appropriate values for the KNN density parameter K and the overlap threshold σ. Optimal values of K were determined through a sensitivity analysis by monitoring the number of initially identified core topic terms against K. The corresponding plot is presented in Figure 3.

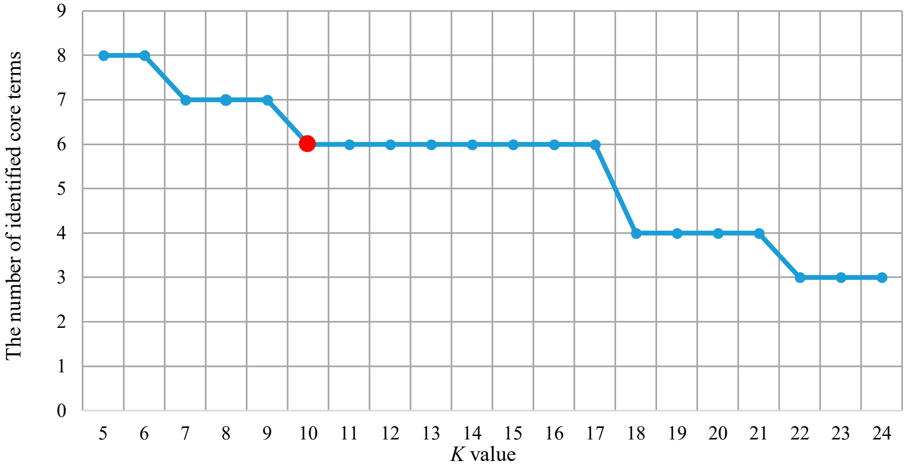


Figure 3. The plot of *K* against the number of identified core topic terms

HTT returned six initial core topic terms at every setting of K between 10 and 17. Therefore, to detect as many topics as possible, we set K to 10 and the overlap threshold σ to 0.8.

Tree generation

With the co-occurrence network as input to the DPS and OCA algorithms, the graph was recursively partitioned into subnetworks of topics in different layers, and the overlaps between topics were evaluated and assigned accordingly. The algorithms stopped at the eighth iteration, yielding a nine-level HTT of computer science research. Figures 4 and 5 illustrate the HTT and detailed terms in topics and their overlaps, respectively.

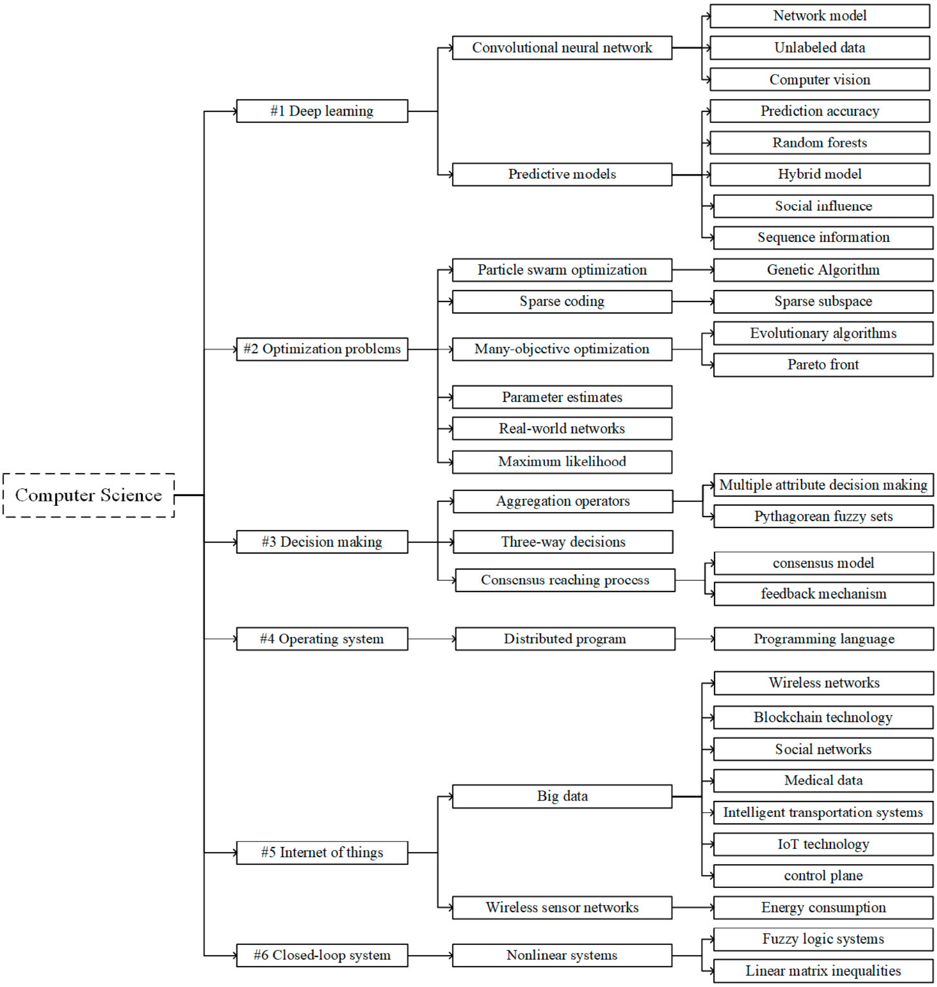
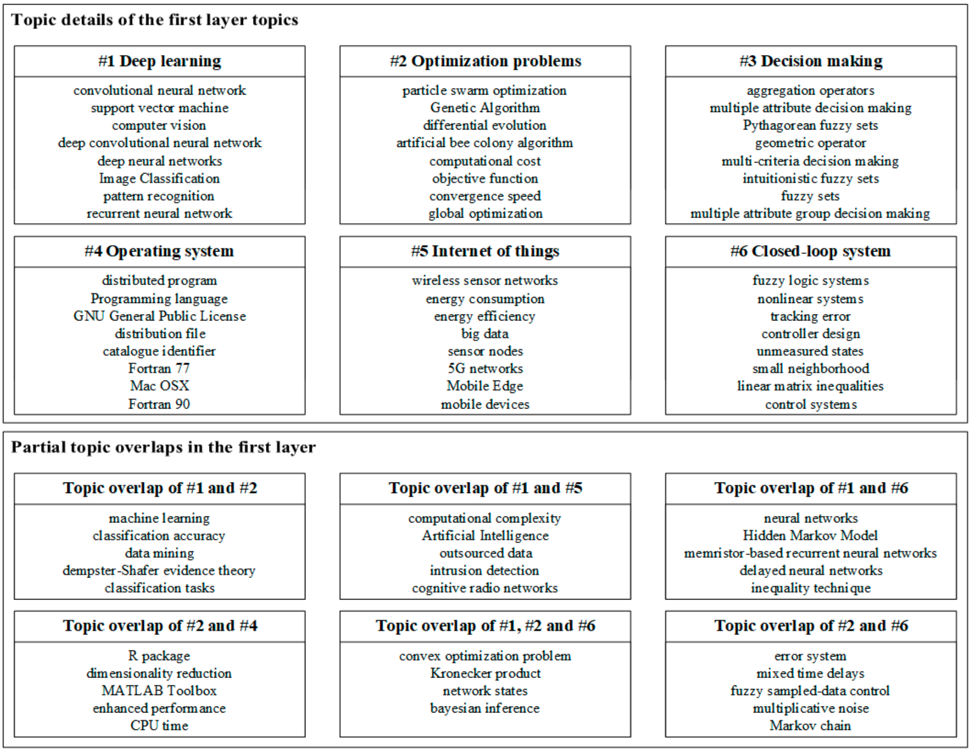


Figure 4. The HTT for computer science[3]

[3] Constraints on the page size limit the tree to its top three layers. The full HTT is available at https://github.com/IntelligentBibliometrics/HTT.



Topic details of the first layer topics

#1 Deep learning
convolutional neural network
support vector machine
computer vision
deep convolutional neural network
deep neural networks
Image Classification
pattern recognition
recurrent neural network

#2 Optimization problems
particle swarm optimization
Genetic Algorithm
differential evolution
artificial bee colony algorithm
computational cost
objective function
convergence speed
global optimization

#3 Decision making
aggregation operators
multiple attribute decision making
Pythagorean fuzzy sets
geometric operator
multi-criteria decision making
intuitionistic fuzzy sets
fuzzy sets
multiple attribute group decision making

#4 Operating system
distributed program
Programming language
GNU General Public License
distribution file
catalogue identifier
Fortran 77
Mac OSX
Fortran 90

#5 Internet of things
wireless sensor networks
energy consumption
energy efficiency
big data
sensor nodes
5G networks
Mobile Edge
mobile devices

#6 Closed-loop system
fuzzy logic systems
nonlinear systems
tracking error
controller design
unmeasured states
small neighborhood
linear matrix inequalities
control systems

Partial topic overlaps in the first layer

Topic overlap of #1 and #2
machine learning
classification accuracy
data mining
dempster-Shafer evidence theory
classification tasks

Topic overlap of #1 and #5
computational complexity
Artificial Intelligence
outsourced data
intrusion detection
cognitive radio networks

Topic overlap of #1 and #6
neural networks
Hidden Markov Model
memristor-based recurrent neural networks
delayed neural networks
inequality technique

Topic overlap of #2 and #4
R package
dimensionality reduction
MATLAB Toolbox
enhanced performance
CPU time

Topic overlap of #1, #2 and #6
convex optimization problem
Kronecker product
network states
bayesian inference

Topic overlap of #2 and #6
error system
mixed time delays
fuzzy sampled-data control
multiplicative noise
Markov chain

Figure 5. The topic details and partial topic overlaps in computer science

Evaluation and discussion

To evaluate the performance of HTT in this case, we calculated the average topic coherence, PCAI, and information loss of the final HTT, with their values presented as 0.619, 0.847, and 6%, respectively. The high PCAI value indicates our methods yield solid and reliable relationships between parent and their corresponding child topics. The low average information loss index suggests that the HTT evenly retains more than 93% of the information in every hierarchy construction process. The topic coherence is above 0.6, which is acceptable in partitioning the tangling research topics in the computer science domain that includes many multi-disciplinary interactions and knowledge convergence.

In Figure 4, the six topics in the first tier reflect six relatively separate research directions, which result from the idea of DPS that each core label should be topologically distant from each other. Simple observation confirms that the selected label terms with high density are also representatives of the terms they lead. Drilling down into each of the six initial parents, *#1 Deep learning* branches off into topics that pertain to various neural network techniques, such as convolutional neural networks and recurrent neural networks, and onwards to the tasks they are used to solve in the real world, e.g., computer vision, image classification, etc. The lower branch of this topic groups the models and metrics associated with deep learning, such as random forest and prediction accuracy. *#2 Optimization problem[s]* spans the different techniques, algorithms, and research objects associated with optimization and its sub-problems. *#3 Decision making* captures the models, strategies, sub-problems relevant to decision intelligence and its processes. #4 *Operating system[s]* groups the research topics surrounding computing architectures and software, which is a fundamental aspect of computer science. #5 *The Internet* *of Things* (IoT) connects big data and sensor technology with its many spheres of application. Last, #6 *Closed-loop system[s]* leads the branch of topics concerning the convergence of computer science with engineering and control systems.

We also generated insights into cross-direction convergence from the topic overlaps in Figure 5. The overlapping terms between #1 and #2 include “data mining”, “classification accuracy”, and “classification tasks”, which are universal concepts for both deep learning and optimization studies. Overlapping terms of #2 and #4 describe two programming tools (R, MATLAB) and computer performance (enhanced performance, CPU time). This overlap indicates a direction of solving optimization problems using computer operating system-based applications. Likewise, the other overlapping terms all indicate different kinds of topic convergence. Intriguingly, “machine learning” was also assigned to this overlapping section. Conventionally, deep learning would be regarded as a sub-topic of machine learning; however, the two terms are close neighbors in this term co-occurrence network, and “deep learning” has a higher KNN density. What this reflects is that deep learning has overshadowed its precursor technologies to become the more dominant research focus. Interestingly, this outcome raises questions over the temporal associations users attach to hierarchies and how the HTT framework prioritizes attention over evolution. This is a question we leave to future study.

Conclusions

This paper presents an end-to-end framework called HTT for identifying topic hierarchies from a co-occurrence network. The methodology combines density peak search and overlapping community allocation to provide a solution that extracts the topics from a corpus, identifies topic overlaps, arranges the topics in a hierarchy, and gives each topic an appropriately descriptive name. In HTT, the core term to each topic in a co-occurrence network, to be used as its label, is determined by the term’s density peak characteristics, while overlapping community allocation detects overlaps among different topics. Recursive implementation of these two algorithms generates a hierarchical topic tree. A case study on the topic hierarchies in computer science demonstrates the feasibility and reliability of the proposed methodology.

In future studies, we plan several improvements to the HTT framework. These include: 1) Automatic parameter tuning: To further improve the adaptability of the methodology, we plan to change the DPS and OCA algorithms into nonparametric functions. Then, optimal values of K and σ could be selected automatically via a maximum entropy model or other approaches. 2) Leveraging additional forms of similarity: Co-occurrence networks are a classical input in bibliometric approaches, but they have also been criticized for their tendency to include too many irrelevant keyword pairs. Other forms of similarity, or combinations of similarities, such as semantic similarity based on topological distance, may prove to be a more effective proxy for the density peak search process. We plan to test these ideas in a future study. 3) HTT for streaming data. We also intend to build a variant of HTT that considers the temporal relationship between topics and how the research topics evolve over time.

Acknowledgments

This work is supported by the Australian Research Council under Discovery Early Career Researcher Award DE190100994.

References

Ba, Z., Cao, Y., Mao, J., & Li, G. (2019). A hierarchical approach to analyzing knowledge integration between two fields—a case study on medical informatics and computer science. *Scientometrics, 119*(3), 1455-1486.

Bai, X., Yang, P., & Shi, X. (2017). An overlapping community detection algorithm based on density peaks. *Neurocomputing, 226*, 7-15.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM), 57*(2), 1-30.

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems, 16*(16), 17-24.

Colavizza, G., & Franceschet, M. (2016). Clustering citation histories in the Physical Review. *Journal of Informetrics, 10*(4), 1037-1051.

Du, M., Ding, S., & Jia, H. (2016). Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems, 99*, 135-145.

Hou, J., Yang, X., & Chen, C. (2018). Emerging trends and new developments in information science: A document co-citation analysis (2009–2016). *Scientometrics, 115*(2), 869-892.

Porter, A. L., Zhang, Y., Huang, Y., & Wu, M. (2020). Tracking and Mining the COVID-19 Research Literature. *Frontiers in Research Metrics and Analytics, 5*, 12.

Qian, Y., Liu, Y., & Sheng, Q. Z. (2020). Understanding hierarchical structural evolution in a scientific discipline: A case study of artificial intelligence. *Journal of Informetrics, 14*(3), 101047.

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science, 344*(6191), 1492-1496.

Shang, J., Zhang, X., Liu, L., Li, S., & Han, J. (2020). *Nettaxo: Automated topic taxonomy construction from text-rich network.* Paper presented at the Proceedings of the Web Conference 2020.

Song, J., Huang, Y., Qi, X., Li, Y., Li, F., Fu, K., et al. (2016). Discovering hierarchical topic evolution in time‐stamped documents. *Journal of the Association for Information Science and Technology, 67*(4), 915-927.

Wang, C., Danilevsky, M., Desai, N., Zhang, Y., Nguyen, P., Taula, T., et al. (2013). *A phrase mining framework for recursive construction of a topical hierarchy.* Paper presented at the Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Wang, C., Liu, J., Desai, N., Danilevsky, M., & Han, J. (2015). Constructing topical hierarchies in heterogeneous information networks. *Knowledge and Information Systems, 44*(3), 529-558.

Wong, W., Liu, W., & Bennamoun, M. (2012). Ontology learning from text: A look back and into the future. *ACM computing surveys (CSUR), 44*(4), 1-36.

Wu, M., Zhang, Y., Zhang, G., & Lu, J. (2020). Exploring the genetic basis of diseases through a heterogeneous bibliometric network: A methodology and case study. *Technological Forecasting and Social Change, 164*, 120513.

Xu, Y., Yin, J., Huang, J., & Yin, Y. (2018). Hierarchical topic modeling with automatic knowledge mining. *Expert Systems with Applications, 103*, 106-117.

Yang, S., Zou, L., Wang, Z., Yan, J., & Wen, J.-R. (2017). *Efficiently answering technical questions—a knowledge graph approach.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., et al. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*(4), 1099-1117.

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change, 85*, 26-39.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology, 68*(8), 1925-1939.

Comparison of citation-based clustering and topic modeling: The case of cardiovascular disease research

Qianqian Xie[1], Ismael Rafols[2], Ludo Waltman[3] and Alfredo Yegros[4]

[1] *q.xie@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University (The Netherlands)

[2] *i.rafols@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University (The Netherlands)
Science Policy Research Unit (SPRU), University of Sussex (England)

[3] *waltmanlr@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University (The Netherlands)

[4] *a.yegros@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University (The Netherlands)

Abstract

We examine the ability of different approaches to discern the topics and the cognitive structure of research into cardiovascular disease (CVD). Since our work is still in progress, we consider only two approaches, namely a citation-based journal map and a topic model, to create maps of CVD research. We compare the results and detect commonalities and differences. We discuss the advantages and disadvantages of these methods and explore reasons behind different results. The results show that the journal map and the topic model generate a relatively similar overall structure of CVD research. The journal map provides an easily understandable intellectual structure of CVD research, while the topic model provides more detailed information on topics and structure. These findings suggest that these two mapping techniques are complementary to a certain extent. The journal map does not suffer from the problem of synonymy and polysemy brought by topic modelling. Meanwhile, the topic model provides more detailed information than the journal map. Also, these findings point to the suitability of both approaches. The journal map is suitable for describing the overall structure of CVD research, and the topic model is suitable for representing more detailed information of topics of CVD research.

Introduction

Science mapping is often used to inform science policy about developments in a research field. Different methodologies for mapping have been developed (Petrovich, 2020). These mapping methods involve different data models and clustering algorithms, and each mapping method exhibits certain advantages and disadvantages, these properties of approach determine different characteristics of the intellectual structure of a research domain (Velden et al., 2017, Chang et al., 2015). Therefore, a concern emerged and is steadily growing: to what extent are results determined by properties of the knowledge structure or by the approaches used (Glaser et al., 2017)? In response to this question, some researchers proposed presenting approaches and comparisons that can address key unresolved theoretical and methodological issues (Glaser et al., 2017). Given that different meaningful and legitimate perspectives on a result produced by a specific approach are possible, rather than aiming at creating the 'best' solution with the highest accuracy, we compare two disparate methodological approaches (Boyack, & Klavans, 2010, Klavans, & Boyack, 2017). There have been few comparisons among approaches. The few studies conducted have shown some similarities yet also significant differences across methods (Velden et al., 2017). However, two of the most widely used methods, namely citation-based clustering and topic model, have not yet been compared. Also, there is a lack of discussion on the advantages and disadvantages of approaches, and addressing this gap would help to get a comprehensive understanding of the suitability of each approach.

In this work, our analysis aims at providing researchers with a better understanding of the strengths and downsides of different approaches. We are particularly interested in how the information obtained from these methods compare to each other, what aspects of results are similar or different, and the reasons behind the possible differences. We propose to use several bibliometric methods to describe the topics and the cognitive structure of a research domain, then compare and detect in what aspects results agree or differ. Next, we discuss the advantages and disadvantages of these methods and explore reasons behind different results.

In this study, we describe the topics and the cognitive structure of the research on CVD, following the delineation of the field provided by Gal et al. (2019), this delineation was created by an expert in cardiology, Prof. Sipido, and collaborators. CVD is an area in which there is an ongoing debate about research priorities and thus where an understanding of intellectual structure is essential. Many stakeholders believe current research is too focused on clinical and biomedical approaches, at the expense of more public health, prevention, and social determinants. The main contribution of our study is that it can answer the question to what extent results are determined by properties of knowledge structure and approaches we use (Glaser et al., 2017). It also contributes to knowledge on the suitability of data models and algorithms for the identification of topics.

Since our research is still in progress, in this study, we consider only two methods, namely citation-based journal map and topic modelling. We characterized the mapping obtained using these two methods, in terms of the content of mapping and the topology of the mapping. As the next step in our project, we plan to also consider mappings generated by other approaches, such as article-level clustering maps, term maps, and maps of MeSH descriptors.

Data

Data was collected from the Web of Science (WoS) database. We used the in-house version of the WoS database at the Centre for Science and Technology Studies (CWTS) at Leiden University. The dataset was established using a hybrid information retrieval strategy, combining lexical and citation-based methods, adopted from Gal et al. (2015). The dataset includes articles, letters, notes, and reviews from 1991 to 2019. It contains 1,079,177 documents from 7,159 journals. The data was retrieved in July 2019.

Methods

Journal map

We applied VOSviewer to build a journal map based on bibliographic coupling links between journals. We collected 7,159 journals from 1991 to 2019 and then removed journals of the subject category 'Multidisciplinary science '. 82 journals were removed, e.g., ‘PLOS ONE’, ‘Scientific Reports’, etc. In the end, we got 7,077 journals. Next, we selected the 400 journals with the largest number of CVD publications. These journals contain 73% of all publications included in the dataset. We then calculated their bibliographic coupling data for constructing a journal map. We removed some general journals, such as ‘Circulation’, ‘American Journal of Cardiology’, etc. Because of their general nature, the inclusion of these journals in a journal map does not contribute to the understanding of the intellectual structure of CVD research.

We aim to compare insights obtained from different approaches, and for this purpose, we prefer to have a relatively fine-grained clustering of the journals in the journal map. Therefore by comparing results produced by choosing different values for the resolution parameter of the clustering techniques used by VOSviewer, we finally chose a value of 1.5 for the resolution parameter. The final map can be seen in Fig.1A. (For map with full detail, please visit https://bit.ly/3gASryG.)

Topic modelling

We trained the Latent Dirichlet Allocation (LDA) topic model contained in the Gensim package to process data. Due to the large-scale final data set, we randomly sampled 10% of the publications with100,286 records. Before training the LDA model, text data was pre-processed. The natural language processing tools were applied to pre-process data. After pre-processing text data, we obtained a dictionary containing 126,784 unique nouns and noun phrases. The LDA model parameters were set to α=50/K, β=0.1 (the default value of β), and Gibbs sampling was run for 2000 iterations in the analysis of the corpus. K indicates the number of topics, and α and β indicate the hyperparameters in the LDA model. K needs to be determined by users, in order to select a suitable number of topics, we repeated experiments of setting the different values of K. We found that terms had a higher probability and better discrimination for K=15. Labels could be obtained by selecting a few meaningful terms among the most frequent (or relevant) terms within every topic. Fig. 1B shows the visualization result of LDA. (If you are interested in the mapping with full detail, please go to https://bit.ly/3rn0h1c to download a file and open the file in your web browser.)

Results

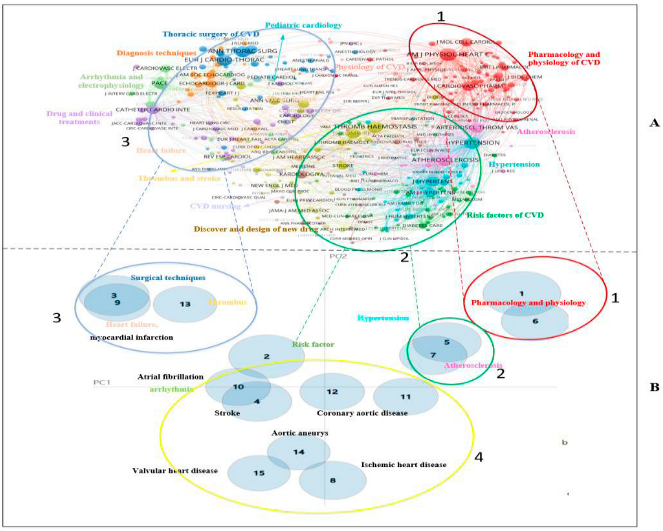


Figure 1. Comparison of journal map and topic modelling map

In Fig. 1A, each journal is represented by a circle, and the distance between two journals approximately represents the relatedness of the journals in terms of bibliographic coupling links. Colors represent clusters of strongly related journals. In Fig. 1B, every circle represents one topic. The distance between the two circles approximately represents the relatedness between topics. We interpreted and compared the map from two aspects: 1) the content of every cluster and topic, 2) the topology of the map.

The research content of clusters and topics

As shown in Fig. 1A, there are 16 specific clusters related to CVD. By interpreting journals within every cluster, we identified various aspects of CVD research: pharmacology and physiology of CVD (cluster 1, red), risk factors of CVD (cluster 2, green), thoracic surgery of CVD (cluster 3, blue), thrombosis and stroke (cluster 4, yellow), drug and clinical treatments (cluster 5, purple), hypertension (cluster 6, blue-green), diagnosis techniques including CT and echocardiogram (cluster 7, orange), discovery and design of new drug (cluster 8, brown), atherosclerosis (cluster 9, pink), physiology of CVD (cluster 10, light red), arrhythmia and electrophysiology (cluster 11, light green), rheumatic heart disease (cluster 12, light blue), CVD nursing (cluster 14, light purple), pediatric cardiology (cluster 15, light blue-green), and heart failure (cluster 16, light orange).

As shown in Fig. 1B, there are 15 specific topics related to CVD. By interpreting terms within every topic, we identified various issues of CVD research. In Pharmacology and physiology (topic 1 and topic 6), research on cell signalling, gene transcription, and animal models are prominent. Risk factor (topic 2) mainly focuses on smoking, diabetes, obesity, BMI, and alcohol. Surgical techniques (topic 3), such as heart bypass, coronary artery bypass graft, and percutaneous transluminal coronary angioplasty. Stroke and its prevention measures (topic 4), with research on life quality and drug prevention. Hypertension (topic 5). Atherosclerosis (topic 7). Ischemic heart disease (topic 8), with research on diagnosis technique and drug prevention. Heart failure and myocardial infarction (topic 9), with research on diagnosing these diseases. Atrial fibrillation and arrhythmia (topic 10) the topic mainly focuses on diagnosis technique and surgery measure. Coronary aortic disease (topic 12). Treatment, therapy, and drug of thrombus (topic 13). Aortic aneurysm (topic 14), which pays attention to the diagnosis technique and surgery measure. Valvular heart disease (topic 15), with research on surgery measure and diagnosis technique.

The topology of the journal map and the topic model mapping

As shown in Fig. 1A, from the overall view, the journal map identifies three main different categories of research: 1) basic research, 2) risk factors of CVD/population, and 3) clinical research (Axel et al., 2018, Gal et al., 2019). Basic research is loosely connected to clinical research, which reveals that there is a gap between basic research and clinical research. In the basic research category, cluster 1 and cluster 10 are closely linked. In the risk factors of CVD category, the relationship of clusters 2, 6, and 9 are strongly interrelated. Meanwhile, cluster 4 is close to cluster 9, because thrombus can lead to atherosclerosis. In the clinical research category, Clusters 5, 11, 7, and 3 are closely linked. These clusters mainly focus on surgery and CVD diagnosis techniques.

As shown in Fig. 1B, from the overall view, the topic modelling map identifies four groups. According to the classification of topics of Gal et al (2019), group 4 can be classified into the clinical research area, which talks about the diseases of CVD. The topic modelling map reveals a similar pattern as the journal map. In basic research, topic 1 and topic 6 are closely linked, and these two clusters are loosely connected to other topics. In clinical research, the relationship of topic 3, 9, and 13, is close. These clusters pay attention to CVD diseases and surgical and drug treatment of these diseases. In risk factors of CVD, Topic 5 and 7 are closely connected, mainly focus on hypertension and atherosclerosis. Topic 2 is loosely connected to topics 5 and 7, while topic 2, 4, and 10 are closely connected.

Comparison between journal map and topic modelling map

Through the comparative analysis of different results produced by different methods, we can get a deeper understanding of the ways in which specific properties of the approaches we used to create differences between our results (Glaser et al., 2017).

In terms of content, the journal map method and topic modelling method both identified the pharmacology and physiology of CVD, risk factors, surgical techniques, and some main diseases of CVD. However, there are also differences, the topic modelling discerned more risk factors, surgical techniques, and CVD diseases than the journal map. Because topic modelling map is created based on document-level data rather than journal-level data. Also, by contrast, text data are probably clearer to users than the bibliographic references data that identified clusters of same content (Zitt et al., 2011). In addition, the journal map may identify some groups that do not appear in the topic model, like the discovery and design of new drug, CVD nursing. Because terms within these clusters are scattered among different CVD disease topics. The reason behind the phenomenon is that terms remain universal on linguistic factors, not confined to scientific databases. While citation-based journal map may prefer to cite the journals from the same corpus (Zitt, 2015).

In terms of topology, both maps show that pharmacology and physiology of CVD are far from other clusters or topics, especially, from surgical treatments. This distance suggests that there are gaps between basic science and clinical research. We discovered some differences as well. From the overall view of these two maps, the journal map provides a clear overall intellectual structure of CVD. Because the intellectual structure of research can be directly read from the organization of bibliographic references. While text data associations, even in full text, do not usually make the knowledge flow apparent due to the polysemy, synonymy used. (Zitt et al., 2011, Zitt, 2015). In the journal map, hypertension, atherosclerosis, and risk factor have a strong connection. However, the risk factor is far from hypertension and atherosclerosis in the topic modelling map. The reasons for this phenomenon may be that textual information can indicate similarities that are not visible to bibliographic reference, and vice versa (Janssens et al., 2008).

Discussion and conclusion

This paper presents the preliminary results of an ongoing project. We created a citation-based journal map and a topic model map to analyze the intellectual structure of CVD research. We compared these two maps from their content and topology. And detected in what aspects results agree or differ and explored reasons behind different results. We found three significant results regarding comparing different maps:

- These two maps may identify relatively similar clusters or topics. However, the topic model map provides more detailed information than the journal map. One reason is that the topic model is created based on document-level data, rather than journal-level data. Another reason is that text data is probably clearer to users than the bibliographic reference data that can identify clusters with the same content (Zitt et al. 2010). That means the topics identified through topic model provide more detailed information that can be used to get a deeper understanding of clusters found on the basis of a journal map.
- The journal map may identify some groups that do not appear in the topic model. Because in topic modelling terms within these clusters are scattered among different CVD disease topics. The reason behind the phenomenon is that terms remain universal on linguistic factors, not confined to scientific databases. While citation-based journal map may prefer to cite the journals from the same corpus (Zitt, 2015).
- These two maps show a gap between basic science and clinical research. Also, the journal map shows an easier to understand overview structure of CVD research because bibliographic references appear as tracers of intellectual connections, the intellectual structure of research can be directly read from the organization of bibliographic references.

Through comparing these two maps, we have a better understanding of the properties of the methods, which can help us learn the suitability of the methods. Topic model processes text data, however, the occurrence of synonymy, polysemy, and common terms may cause problems and may make it difficult to interpret topics' meaning. Therefore require complex semantic processing and pre-processing. By contrast, the journal map is a citation-based method. It does not need too complicated processing, and it can provide a clear intellectual structure of research. However, the journal's bibliographic references provide less detailed information. Our work shows that these two methods are complementary to a certain extent. The journal map describes the intellectual structure of CVD research in a way that is easy to understand, and it does not suffer from the problems of synonymy and polysemy. At the same time, the topic model provides more detailed information, which makes up for the limitations of the citation-based journal map.

This research is still in progress, and therefore there are a lot of possibilities for improving and extending the research. In this paper, we used only two methods to map CVD research. In the future, we plan to add more approaches, including maps based on article-level clustering, MeSH descriptors, and terms. These methods could be combined with our journal map and topic model, which may provide a richer understanding of the CVD field and of the properties of the different methods.

Acknowledgments

We thank Diane Gal and Karen Sipido for kindly helping us in using the search strings for the delineation developed in Gal et al. (2019). Qianqian Xie is financially supported by the China Scholarship Council (Grant No. 201906540033).

References

Axel, R.P, Anastasis, N., et al. (2018). CardioScape mapping the cardiovascular funding landscape in Europe. *European Heart Journal*, 39, 2423-2430.

Boyack, K.W. & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12), 2389–2404.

Chang, Y.W., Huang, M.H. & Lin, C.W. (2015). Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. *Scientometrics*, 105, 2071–2087.

Gal, D., Sipido K.R. & Glänzel, W. (2015). Using bibliometrics-aided retrieval to delineate the field of cardiovascular research. *Proceedings of ISSI 2015 Istanbul Istanbul: Bogaziçi University Printhouse*, 1018–1023.

Gal, D., Thijs, B., Glänzel, W. & Sipido, K.R. (2019). Hot topics and trends in cardiovascular research. *European Heart Journal*, 40(28), 2363-2374.

Glänzel, W. & Thijs, B. (2017). Using hybrid methods and 'core documents' for the representation of clusters and topics: The astronomy dataset. *Scientometrics*, 111(2), 1071-1087.

Gläser, J., Glänzel, W. & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, 111, 981–998.

Janssens, F., Glänzel, W. & De Moor, B. (2008). A hybrid mapping of information science . *Scientometrics*, 75(3), 607-631.

Klavans, R. & Boyack, K.W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, 68(4), 984–998.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy*, 18(4), 209-223.

Petrovich, E. (2020). Science mapping. *ISKO Encyclopedia of Knowledge Organization*. www.isko.org/cyclo/science_mapping.

Velden, T., Boyack, K.W., Gläser, J., Koopman, R., Scharnhorst, A. & Wang, S. (2017). Comparison of topic extraction approaches and their results. *Scientometrics*, 111(2), 1169-1221.

Zitt, M. (2015). Meso-level retrieval: IR-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation. *Scientometrics*, 102(3), 2223-2245.

Zitt, M., Lelu, A. & Bassecoulard Zitt, E. (2011). Hybrid citation-word representations in science mapping: Portolan charts of research fields? *Journal of the American Society for Information Science and Technology*, 62(1), 19-39.

Interdisciplinarity vs. Unidisciplinarity: A Structural Comparison of Multi-generation Citations and References

Jiawei Xu[1], Chao Min[2], Win-bin Huang[1] and Yi Bu[1]

[1]*{imxujiawei, huangwb, buyi}@pku.edu.cn*
Department of Information Management, Peking University (China)

[2]*mc@nju.edu.cn*
School of Information Management, Nanjing University (China)

Abstract

This paper compares the structural characteristics between the citation/reference cascades (networks containing multi-generation citations and references) of interdisciplinary and unidisciplinary papers. By constructing multiple citation/reference cascades for each publication in the American Physical Society (APS) dataset, we found that: (1) for papers having a similar number of references, interdisciplinary papers tend to have a smaller and less 'radioactive' knowledge foundation than unidisciplinary papers, though the knowledge foundation of interdisciplinary papers is more heterogeneous; (2) for papers having a similar 'direct' scientific impact, unidisciplinary papers have a relatively greater 'indirect' impact, which indicates that the scientific impact of unidisciplinary papers is 'deeper' and more persistent; and, (3) compared to unidisciplinary papers, there exists at least a short time in which interdisciplinary papers trigger more follow-up discussions.

Introduction

Interdisciplinary research is found to facilitate knowledge integration and scientific innovation (Klein, 1990; Rafols et al., 2012). Yet, some researchers also hold the view that "the general superiority of interdisciplinarity over disciplinary knowledge" has not been fully proved (Jacobs & Frickel, 2009, p. 60). Within this debate, a considerable number of studies have focused on interdisciplinarity, many of which have paid special attention to the relationship between the interdisciplinarity of a single publication and its scientific impact. For instance, Larivière and Gingras (2010) found that publications in physics citing citation-intensive disciplines tend to obtain higher citation scores. Chen, Clément Arsenault and Larivière (2015) showed that the top 1% most-cited papers present higher levels of interdisciplinarity. These studies suggest to us that there might be a positive correlation between the level of interdisciplinarity of a publication and its 'direct' scientific impact, calculated by citation count-related indicators. Nonetheless, for two publications having a similar 'direct' impact, one might have a greater 'indirect' impact (quantified by using multi-generation citing publications) than the other. And the publication that triggers more follow-up discussions seems more invaluable.

To this end, in this paper, we adopt an existing citation network structure, namely citation cascade (Min, Sun & Ding, 2017; Min et al., 2021a) and reference cascade (Min et al., 2021b). For a given focal publication, its citation cascade contains its first-, second- (i.e., citing publications' citing publications), …, and *n*th generation citing publications, as well as their citing relationships. Symmetrically, its reference cascade includes its first-, second-, …, and *n*th generation references. Rousseau identified the concept of multi-generation citations and references dating back to 1987 (Rousseau, 1987), a concept which has inspired many further studies (Hu, Rousseau & Chen, 2011; Hu & Rousseau, 2016). Investigating citation cascades helps understand forward citation behaviour details and how a particular study inspires later research, regardless of whether this influence is direct or indirect. As for reference cascades, they include backward citation nuances in terms of knowledge foundations (e.g., scale, interlock-wise structure, etc.). This paper employs these two cascades and compares the 'indirect' scientific impact and knowledge foundation of interdisciplinary and unidisciplinary publications by adopting many structural-level measurements, such as structural virality, average clustering coefficient and network density.

Data and methods

Data

We adopt the American Physical Society (APS) dataset that covers 541,448 publications from 1893 to 2013 and the nearly 3 million citing relations among them. As a balance to guarantee a relatively long citation window and a sufficient number of publications, we particularly select all publications in 1996, 1997 and 1998 in our following empirical study.

All publications in or after 1975 are assigned one or more codes within the Physics and Astronomy Classification Scheme (PACS). These codes are adopted to identify fields in physics. There are several levels in the PACS. For simplicity, we employ the top-level codes (representing 10 fields) in our study to present the specialties of publications.

Methods

Citation cascade and reference cascade

From the structural perspective, citation cascade and reference cascade are generated through citing relations. For a given focal publication, its citation cascade comprises its citing publications (first-generation citations), its citing publications' citing publications (second-generation citations), …, and *n*th-generation citations. A reference cascade is symmetric; that being said, the reference cascade of a given publication contains its references (first-generation references), its references' references (second-generation references), …, and *n*th-generation references. Thus, reference and citation cascades tend to be more informative than traditional bibliometric indicators, e.g., number of citations and its normalised variants (Bornmann & Daniel, 2008).

Previous scientometricians have focused on the conceptualisation and operationalisation of citation cascades (e.g., Min et al., 2017; Min et al., 2021a), described as "the constitution of a series of subsequent citing events initiated by a certain publication" (Min et al., 2021a, p. 110). Similarly, reference cascades contain a series of proceeding citing (referencing) events initiated by a certain publication.

We defined all publications in 1996, 1997 and 1998 as the 'initial publication set' ***I***. And, for every publication in ***I***, we created its citation cascade and reference cascade using all the citing relations in the APS dataset. We stipulate that, if the depth of a publication's reference or citation cascade is less than four, this publication, as well as its cascades, will be removed from our following empirical study. To this end, 29,310 publications remain.

Quantifying interdisciplinarity

As mentioned above, each publication has one or more PACS labels from the 10 labels. Following Porter and Chubin (1985), the references' PACS labels are used to measure how interdisciplinary a publication is. We characterise the degree of interdisciplinarity of one specific publication following Figure 1. As shown, we first examine the PACS labels of all references of i and adopt the Reverse Simpson Index (RSI_i) as an indicator:

$$RSI_i = 1 - \sum_{j=1}^{R} {p_j}^2$$

where R represents the number of fields publication i's references come from and p_j the proportion of field j among all fields. Per Simpson (1949), RSI is "a measure of the concentration of the classification" (p. 688). The greater the RSI is, the higher the degree of interdisciplinarity. In practice, we define the top 20% ranked publications in terms of RSI as interdisciplinary papers and the bottom 20% as unidisciplinary papers[1]. This results in 5837 and 5834 publications in each group, respectively[2].

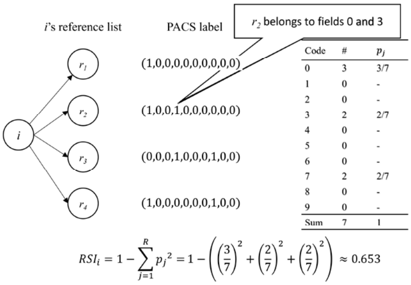


Figure 1. Illustration of quantifying interdisciplinarity of a publication *i*.

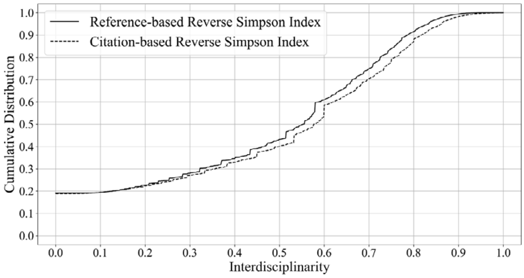


Figure 2. The cumulative distribution of the two interdisciplinarity indicators.

Matching

To reveal the differences regarding interdisciplinarity and unidisciplinarity, we set up two groups of publications. For each unidisciplinary publication, we try to find a 'similar' publication in the interdisciplinary group. To this end, we employ a one-to-one matching and stipulate that one publication cannot be matched with multiple publications. When implementing this matching process, we follow the below rules:

(1) This pair of publications have at least one same PACS field;
(2) For a given unidisciplinary publication, we only keep a certain interdisciplinary publication that has >90% field similarity[3];
(3) Compared with the higher cited publication (regardless of which groups), the lower cited publication has at least 90% as many citations, i.e., their numbers of citations have a ≤10% difference;
(4) Compared with the higher citing publication (regardless of which groups), the lower citing publication has at least 90% as many as references;
(5) This pair of publications have the same document type; and
(6) This pair of papers were published in the same journal in the same year.

Finally, we obtained 822 pairs of publications. The matched groups (822*2=1644 publications) are hereafter abbreviated as interdisciplinarity and unidisciplinary groups.

[1] According to Figure 2, there are ~20% publications with an *RSI* of 0.0. We therefore set the bottom 20% as unidisciplinary publications. Correspondingly, the top 20% are defined as interdisciplinary ones.

[2] Because there might be more than one publication exactly at the 20% threshold, we have to include all of these publications in practice. This is why there are minor differences regarding the number of publications in the two groups.

[3] For each publication, we construct a 10-dimension vector that contains its PACS code; each dimension represents one of the 10 fields under physics. If a publication has a certain field label, this dimension will be marked as one; otherwise, zero. We compare two publications' field similarity by calculating the cosine similarity of the two vectors.

Measurements

For each publication in the unidisciplinarity or interdisciplinarity group, we construct three citation cascades with different generations, i.e., citation cascade with the first two, three and four generations, as well as three reference cascades with different generations, i.e., reference cascade with the first two, three and four generations. We calculate the below indicators particularly:

- Structural popularity (SP): For citation cascades, SP represents the range to which knowledge is diffused; in terms of reference cascades, it indicates the scale of knowledge foundation. We utilise the number of nodes in a cascade as a measurement.
- Structural virality (SV): SV measures the knowledge diffusion depth of the publication. We use the average depth of the publication's citation cascade to represent it.

$$SV = \frac{1}{|T|-1} \sum_{v \in T, v \neq i} d(v, i),$$

T is the cascade initiated by publication i, $|T|$ is the number of nodes in T and $d(v, i)$ is the shortest path between v and i.

- Average clustering coefficient (ACC): ACC reflects the goodness of clustering in a network. Specifically:

$$ACC = \frac{1}{|V|} \sum_{v_k} C(v_k)$$

where V is the node set and $C(v_k)$ is the clustering coefficient of node v_k:

$$C(v_k) = \frac{number\ of\ closed\ triads\ connected\ to\ v_k}{number\ of\ triples\ of\ vertices\ centered\ on\ v_k}$$

- Network density (ND): ND describes the proportion of the potential connections in a network that are actual connections.

$$ND = \frac{2|E|}{|V|(|V|-1)}$$

where E is the edge set. The higher the ND is, the more connections the network has.

Results and Discussion

To statistically test whether the interdisciplinary and unidisciplinary publications have significant differences regarding the aforementioned indicators, we employ the Wilcoxon signed-rank test. As shown in Table 1, in terms of reference cascades, papers from the two groups have shown significant differences in SP and ACC when we select the first three and four generations in the cascade. As for citation cascades, we observe a significant difference for SV and ND for the first two and three generations. We did not find any significance in the first four generations for citation cascades, partly because of, at least conceptually, the insufficient topical similarity between a focal publication and its fourth-generation citing publications.

Table 1. Wilcoxon signed-rank test results.

Indicator	SP	SV	ACC	ND
Reference Cascade (first 2 generations) Sig.	0.149	-	0.240	0.379
Reference Cascade (first 3 generations) Sig.	**0.027***	-	**0.000*****	0.652
Reference Cascade (first 4 generations) Sig.	**0.005****	-	**0.000*****	0.406
Citation Cascade (first 2 generations) Sig.	0.108	**0.007****	0.269	**0.048***
Citation Cascade (first 3 generations) Sig.	0.105	**0.036***	0.076	0.069
Citation Cascade (first 4 generations) Sig.	0.208	0.117	0.198	0.139

Notes. SP=Structural popularity; **SV**=Structural virality; **ACC**=Average clustering coefficient; **ND**=Network density. $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$. All significant results under the 0.05 threshold are **bolded**.

We are particularly interested in the indicators that show statistically significant differences. In this research-in-progress paper, we present four of them in Figure 3, representing SP and ACC for reference cascades (sub-figures A and B, respectively) and SV and ND for citation cascades (sub-figures C and D, respectively). Figure 3(A) indicates the cumulative distribution of reference cascade SP (first four generations). We can see that, when SP<400, the unidisciplinarity and interdisciplinarity groups distribute closely and that there are more interdisciplinarity papers whose SPs are between 400 and 1100. Yet, when SP is greater, unidisciplinary papers dominate. As SP of reference cascades represents the knowledge foundation of the focal publication, Figure 3(A) reveals that interdisciplinary publications tend to have a smaller amount of knowledge foundation than unidisciplinary publications.

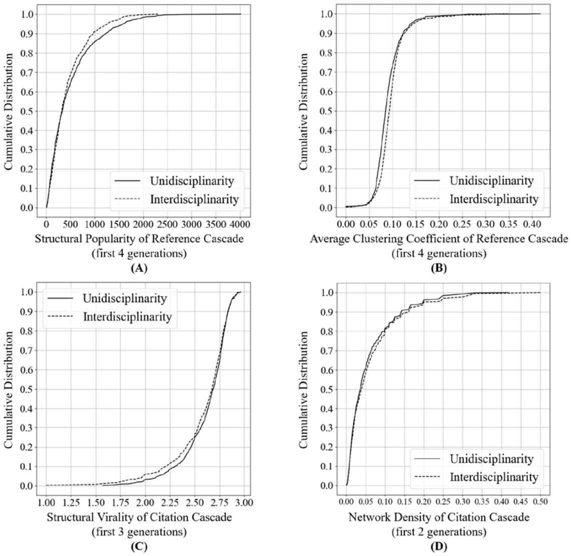


Figure 3. The cumulative distribution of the four structural indicators.

Figure 3(B) shows the cumulative distribution of reference cascade ACC (first four generations), and it demonstrates that an interdisciplinary paper's reference cascade is more similar to an interlocked network, which reveals that an interdisciplinary paper's knowledge foundation is more 'interlocked', though interdisciplinarity reflects the diversity of references.

As to citation cascade, Figure 3(C) presents the cumulative distribution of citation cascade SV (first three generations), in which we find that there are more interdisciplinary publications with a lower SV value. As SV is quantified by the average 'distance' between the focal publication and other publications in the citation cascade, this demonstrates that interdisciplinary publications have a more 'direct' or 'shallow' impact while unidisciplinary publications have a more 'indirect' or 'deep' impact.

Nonetheless, citation cascade ND (first two generations) displays the opposite pattern – we observe more interdisciplinary publications that have a greater value of ND (Figure 3(D)). Figure 3(D) reveals that there are more connections between the publications in the citation cascade of an interdisciplinary paper, and that interdisciplinary research may intrigue researchers and results in more discussions.

Conclusions

In this paper, we implement a structural comparison between the reference/citation cascades of unidisciplinary and interdisciplinary papers. With a sophisticated matching process, our study suggests that interdisciplinary papers tend to have a smaller knowledge foundation and a more 'shallow' or 'direct' impact on the follow-up research. Nonetheless, interdisciplinary papers contribute to the interactions between researchers. We speculate that interdisciplinary research may be relatively pioneering and there are fewer prior works for researchers to build on. Due to this characteristic of interdisciplinary research, policy makers should even more strongly encourage interdisciplinary research and devote more resources to it. Also, more future works are needed to explore the features and mechanisms of interdisciplinary research.

There are some limitations in our study. First, we do not set up any restriction when constructing the cascades, e.g., publication year of forward citations. Future work may consider involving a temporal dimension as a filter and set up different thresholds to examine differences. Second, since our experiment is implemented only in the discipline of physics, we cannot generalise the conclusions to other domains. In the future, we are going to apply the current empirical study in a more comprehensive, cross-discipline bibliographic dataset. Finally, only a few constraints are taken into account in our matching experiment, which may influence the reliability of the experiment. Many other factors could be controlled in our future study, such as the number of co-authors in a publication and their affiliations.

Acknowledgments

We acknowledge the two anonymous referees as well as the organisers for their useful comments. We also acknowledge the American Physical Society (APS) for publicly share the dataset for research.

References

Bornmann, L. & Daniel, H. D. (2008). What do citation counts measure? a review of studies on citing behaviour. *Journal of Documentation, 64*(1), 45-80.

Chen, S., Arsenault, C. & Larivière, V. (2015). Are top-cited papers more interdisciplinary? *Journal of Informetrics, 9*(4), 1034–1046.

Hu, X. & Rousseau, R. (2016). Scientific influence is not always visible: the phenomenon of under-cited influential publications. *Journal of Informetrics, 10*(4), 1079-1091.

Hu, X., Rousseau, R. & Chen, J. (2011). On the definition of forward and backward citation generations. *Journal of Informetrics, 5*(1), 27–36.

Jacobs, J. A. & Frickel, S. (2009). Interdisciplinarity: A critical assessment. *Annual Review of Sociology*, 35, 43–65. http://dx.doi.org/10.1146/annurev-soc-070308-115954

Klein, J.T. (1990). *Interdisciplinarity: History, theory, and practice.* Detroit, MI: Wayne State University Press.

Larivière, V. & Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology, 61*(1), 126-131.

Min, C., Chen, Q., Yan, E., Bu, Y. & Sun, J. (2021a). Citation cascade and the evolution of topic relevance. *Journal of the Association for Information Science & Technology*, *72*(1), 110-127.

Min, C., Xu, J., Han, T. & Bu, Y. (2021b). References of References: How Far is the Knowledge Ancestry. https://arxiv.org/abs/2101.08577

Min, C., Sun, J. & Ding, Y. (2017). Quantifying the evolution of citation cascades. *Proceedings of the Association for Information Science and Technology, 54*(1),761-763.

Porter, A.L. & Chubin, D.E. (1985). An indicator of cross-disciplinary research. *Scientometrics, 8*(3–4), 161–176

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P. & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy, 41*(7), 1262–1282.

Rousseau, R. (1987). The Gozinto theorem: Using citations to determine influence on a scientific publication. *Scientometrics*, 11, 217–229.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163, 688.

RStarRank: A Method for Identifying Rising Stars based on Academic Publication

Jing Xu[1], Chuan Tang[2], Lu Tang[3]

[1] *jingxu@clas.ac.cn*
Chengdu Library and Information Center, Chinese Academy of Sciences, 16, South Section 2, Yihuan Road, Chengdu, Sichuan 610041 (China)

[2] *tangc@clas.ac.cn*
Chengdu Library and Information Center, Chinese Academy of Sciences, 16, South Section 2, Yihuan Road, Chengdu, Sichuan 610041 (China)
Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy Sciences, Beijing 100049 (China)

[3] *tanglu@clas.ac.cn*
Chengdu Library and Information Center, Chinese Academy of Sciences, 16, South Section 2, Yihuan Road, Chengdu, Sichuan 610041 (China)

Abstract

Identifying rising stars in scientific fields is critical to many kinds of applications. Traditional talent evaluation methods are inapplicable for the recognition of rising stars. Moreover, existing recognition methods demonstrate some problems and limitations. This paper proposes a ranking-based method: RStarRank, which recognizes rising stars with machine learning algorithm. In other words, this method generates a prediction model through learning preliminary features of many beginners and their future H-index to predict sequential probability of young talents' growth into academic leaders. This paper takes the computer science subject for empirical study. The results show that, the prediction accuracy can reach over 90% for 10-year model and over 84% for 5-year model, RStarRank has been verified to have a certain validity and has a higher prediction accuracy than other similar algorithms. Finally, this paper summarizes and describes deficiencies of the proposed method.

Introduction

Young scientists are emerging forces that can advance future scientific development. Identifying scientific rising stars is critical for the recruitment of research institutions, cultivation of youth talents, award assessment, and selection of peer reviewers, etc. However, there are few related studies on this problem at present. Generally, identifying rising stars is an academic evaluation issue. Traditional academic evaluation methods are mainly applicable to evaluate and recognize scientific talents with longer careers, making it difficult to recognize rising stars. Consequently, studying the recognition of rising stars will help to meet practical needs, and promote the evaluation of scientific talents, thus offsetting deficiencies in evaluating and recognizing scientific youth talents.

This paper aims to study rising stars and identifying them among young talents. After summarizing deficiencies of existing studies, this paper provides RStarRank, a new method based on ranking and machine learning algorithm. Then, an empirical analysis is conducted. Finally, this paper summarizes deficiencies of RStarRank and describes outlook for future study.

Related Work

Compared to researchers with longer academic careers, beginners carry some unique characteristics, making it difficult to recognize beginners simply with traditional recognition and evaluation method. Li (2009) firstly stated the issue of recognizing rising stars in 2009. According to the framework on which they are based, they are divided into 3 types in this paper, including ranking-based, classifying-based, and clustering-based method.

Ranking-based method

By establishing indicators and indicator system representing features of rising stars (possibly unknown), the ranking-based methods calculate "comprehensive indicator value" of each beginners and rank them to obtain a set of rising stars. Some ranking algorithms are PubRank, StarRank and CocaRank (Ning, Liu & Kong, 2017) (Ning, Liu, Zhang & Wang, 2017) (Daud, Aljohani, Abbasi, Rafique, Amjad, Dawood & Alyoubi, 2017) (Zhang, Liu, Yu, Zhang & Zhou, 2017) (Wijegunawardana, Mehrotra & Mohan, 2016). (Zhang, Ning, Bai, Wang, Yu & Xia, 2016) (Zhang, Xia, Wang, Bai, Yu, Bekele & Peng, 2016).

Classifying-based method

This method classifies beginners into those who may become highly influential scholars and those who may not. Through learning indicator features of the said two groups, this method generates a prediction model for recognition and prediction. This method mostly uses supervised learning (Billah, 2013) (Daud, Ahmad, Malik, M. S., & Che, D. 2015), to predict rising stars.

Clustering-based method

Tsatsaronis et al (Tsatsaronis, Varlamis, Torge, Reimann, Nørvåg, Schroeder & Zschunke, 2011) (Panagopoulos, Tsatsaronis & Varlamis, 2017) regarded the recognition of rising stars as a clustering issue. They classified scholars into four levels, i.e. well-established, rising stars, stable publication rate, and declining authors. Tsatsaronis et al used unsupervised learning method for scholar clustering, and conducted feature description for the obtained four types based on basic features, thus recognizing rising stars.

Table 1. Advantages and weaknesses of three methods for rising stars identification

Method	*Advantages*	*Weaknesses*
Ranking-based method	Reflects the potential of scholars to become rising stars. Capable of recognizing a number of stars.	The selection of feature indicators is critical. Some algorithms are highly complex due to feature indicator processing.
Classifying-based method	The results are definite, i.e. classifying the beginners into rising stars or those who are not.	The feature indicators and thresholds are particularly critical. It is impractical to simply classify beginners into two types.
Clustering-based method	Beginners are classified into several types.	The clustering results may be inconsistent with this assumption. Therefore, clustering results may not be highly interpreted.

Problem exists

Classifying-based method divides beginners into two categories, the two groups is greatly affected by the selection of the index threshold, thus the identification may fail once the index threshold is selected unscientifically or unreasonably. While clustering-based method relies on the clustering of the characteristics of academic beginners and the clustering results might be contradictory to the assumptions. The ranking-based method assumes that academic newcomers demonstrate different levels of potential and probability to become rising stars, which conforms to the objective law. Therefore, this paper selects ranking-based method to conduct the research. However, problems still exist:

- The definition of the key issues of identifying rising stars do not match the actual application scenario and there is a lack of reasonableness. For the definition of beginners in scientific research, some intercepted all authors for a period of time, some selected the authors that are not in the Top 3%. None of these definitions fit the actual application scenario. For the setting of the "future" observation time, some studies set it as 3-10 years, which is too short to assess their achievements.

- Methods are not universally applicable in different fields. The indicators of scholars' academic achievements vary substantially across fields. For example, citations in the medical field are often much higher than those in other fields.
- These methods require a large amount of data and high computing cost. Most methods require a large data set and complex data processing, such as heterogeneous information networks. Other methods rely on complex data indicators, such as PageRank value, etc.
- The prediction accuracy needs to be improved. The existing accuracy verification of ranking-based method mainly compares the authors' indicators of H-index, number of papers, and frequency of citations. Judging from the scale and effect of the current verification, the verification method and its accuracy are to be improved.

Method for rising star identification: RStarRank

Problem definition

In this study, the identification of academic star is considered to identify the Top K academic beginners most likely to grow into academic rising stars, namely:
For beginners $a_i(a_i \in \mathrm{A}), |\mathrm{A}| = \mathrm{N}$ in academic beginners set A, the academic career of a_i starts at time T1, p_i represents the probability of a_i to become a rising star at time T2, $X_i(X_i \in R^m)$ represents the feature index vector of a_i during time period (T1~T2), m represents the number of eigenvalues, y_i represents the H-index of a_i during time period (T1~T3), and Y represents the H-indexes of all beginners in set A.

This study aims to obtain P' on the basis of X_i using prediction model M by training the machine learning algorithm to learn the correlation between X_i and y_i. Here, P' represents the prediction probability sequence from large to small of all beginners in set A; P represents the probability sequence from large to small of all beginners in set A, which is calculated using Y. This is based on the assumption that the higher the H-index is, the higher the possibility that the scholar will become an academic rising star. In this study, H-index of a scholar in the 30th year of his/her academic career is used as a reference index, reflecting his/her level of research achievements to a certain extent. The corresponding relationship between P and Y is shown in Formula 1.

$$Order_{ij} = \begin{cases} 1, & if\ (y_i > y_j)\text{or}(y_i = y_j \text{ and } SN_i > SN_j), \\ & a_i \text{ ranks in front of } a_j \text{ in sequence P} \\ -1 & if\ (y_i < y_j)\text{or}(y_i = y_j \text{ and } SN_i < SN_j)\ , \\ & a_j \text{ ranks in front of } a_i \text{ in sequence P} \end{cases} \qquad \text{(Formula 1)}$$

For $\forall a_i, a_j \in \mathrm{A}$, $Order_{ij}$ represents the relative position of a_i, a_j in the sequence P, SN_i represents the author number of a_i, $Order'_{ij}$ represents the relative position of a_i, a_j in sequence P'.
This research aims to make P' overlaps with P, i.e., $\forall a_i, a_j \in \mathrm{A}, Order'_{ij} = Order_{ij}$. Therefore, it is needed to maximize the value of $\delta(\mathrm{P}, \mathrm{P}')$.

$$\delta(\mathrm{P}, \mathrm{P}') = \sum_{a_i, a_j \in \mathrm{A}}^{Order'_{ij} = Order_{ij}} Order'_{ij} * Order_{ij} \qquad \text{(Formula 2)}$$

The process of RStarRank is shown in Figure 1.

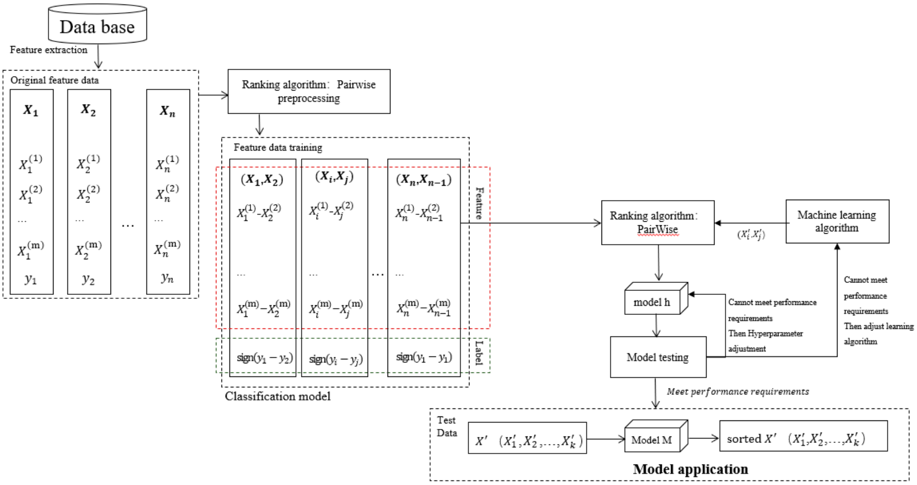


Figure 1. RStarRank Flow Chart

Specific issues involved are more clearly defined as below.

Standards for academic beginners. As rising stars have shorter academic careers, this paper considers the length of academic careers as the criteria to identify beginners. So, beginner a_i should initially publish paper at time T1, i.e. T1 is the starting point of his/her academic career.

Why H-index. This paper use H-index as the standards for highly influential scholars. Although there are certain problems in adopting H index, there are three reasons that support the adoption in this paper. Firstly, H index is more suitable for the summative evaluation of the whole academic career of a scholar. The method validation requires the tracking of a scholar for 20-30 years, which means the scholar should have reached the academic golden age of 45-55 at this time, hence H-index is a more appropriate index to measure his/her academic achievements. Secondly, some studies have shown that although the H-index standards of important scholars in different fields may vary, but H-index, generally speaking, is positively correlated with the academic level of researchers. Some scholars believe that those with H-indexes of 40 or even higher could be regarded as outstanding scientists, and those with H-indexes of 60 or even higher could be regarded as unique scholars (Meho Li, 2007). Therefore, it is reasonable to use H-index to reflect the relative situation of scholars in the academic golden period. Finally, the calculation of H-index is simple and comprehensible, facilitating its promotion and application.

Time periods. The choice of time T2 should consider the time period when the potential of the beginner may be best reflected. This paper selects two time periods to train the samples: the intervals between time T1 and time T2 are 5 and 10 years, and mapping these two time periods into the feature indicators respectively. Since H-index is more suitable to reflect scholars with a long academic history, the choice of time T3 should be able to reflect the real level of scholars more accurately. This paper uses the 30th year in a scholar's academic career as time T3 for two reasons: the 30th year of a researcher's academic career corresponds to his/her biological age of about 60, making it objective to use H-index to judge the level of scientific research achievements at this time; considering the integrity of data collection, the data in the early stage of database construction may have non-standard data format and missing data items, which is not conducive to analysis and modeling.

Prediction algorithm

This paper uses Pairwise ranking algorithm for supervised learning. Two of n beginners from the set of beginners A are combined into a group, i.e. sample is (X_1, X_2),

(X_1, X_3),...(X_i, X_j)...(X_n, X_{n-1}); and h(X_i- X_j) is prediction function of the ranking pair; X_i and X_j are eigenvalues of beginner $a_i$and a_j at time T1-T2; $y_i$and y_j are H-index of beginner $a_i$and $a_j$at time T3. The input of learning model is <X_i, X_j,sign(y_i- y_j)>, where $i \neq j$, $y_i \neq y_j$. sign(x) is a sign function, as shown in Equation 3.

$$\mathrm{sign}(y_i - y_j) = \begin{cases} 1 \quad if\ y_i > y_j, \\ \text{Larger probability of becoming a rising star for } a_i \text{ compared to} a_j \\ -1 \quad if\ y_i < y_j, \\ \text{Smaller probability of becoming a rising star for } a_i \text{ compared to } a_j \end{cases} \qquad \text{(Formula 3)}$$

The prediction target here is to enable h(X_i- X_j) to infinitely approximates sign(y_i- y_j). Therefore, the ranking issue is converted into a binary prediction issue.

Feature selection

The features here are classified into different groups, including author, publication, institution, and time increment. This paper uses 5-year and 10-year data, with feature indicator set R1 and R2 respectively, as shown in Appendix 1. For the publication here, periodical / meeting level proposed by subject matter experts is used to represent paper quality. For the institution, existing and recognized ranking of institutions is used as a basis of institutional grade.

Model testing

As a binary issue, the binary performance index is used for measurement. The final output of the model would be a list of ranking for rising stars. Consequently, ranking-type performance index can be used for measurement as below:

- *Accuracy, precision, recall rate & F1*

$$\text{Accurancy (Accurancy rate)} = \frac{TP + TN}{TP + FP + FN + TN} \qquad \text{(Formula 4)}$$

$$\text{Precision (Precision rate)} = \frac{TP}{TP + FP} \qquad \text{(Formula 5)}$$

$$\text{Recall (Recall rate)} = \frac{TP}{TP + FN} \qquad \text{(Formula 6)}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad \text{(Formula 7)}$$

(TP is the number of positive real values and prediction values; FP is the number of negative real values and positive prediction values; FN is the number of positive real values and negative prediction values; TN is the number of negative real values and prediction values.)

- *Hit Ratio (HR) of Top K ranking*

Among Top K ranking, HR is the probability of hitting prediction value among Top K.

$$\text{HR} = \frac{N\text{umber of Hits}}{N\text{umber of Sample}} \qquad \text{(Formula 8)}$$

Validation

Validation of indicator data. The RStarRank will eventually obtain the ranking list of the probability of beginners becoming rising stars. For the top beginners in the ranking sequence, three indicators H-index, total amount of papers and the number of citations after 30 years (in the later period of their academic career) are investigated. In the later stage of academic career, these indicators are important indicators of scholars' academic influence. Therefore, if the predicted ranking result is better, the indexes of scholars at the top of the predicted ranking sequence should also be higher 30 years later.

Verification of "future" academic achievements. Scholars in the later stages of their academic career can be verified by their academic achievements, including but not limited to the organizations they work for, the honors and the titles they have won.

Comparison of effectiveness of the methods. The existing rank-based methods with similar data size and computation cost to RStarRank are selected for comparison. This study intends to select PubRank method, as this method shares similar algorithm complexity with RStarRank, the calculation amount is also close to the complexity of the late RStarRank application, and the data used is similar to RStarRank method.

As it is the comparison of two ranking prediction results of academic beginners, the ranking comparison is carried out from two aspects. The first is to examine the H-index, total number of papers and the number of citations of academic beginners at the top of the ranking. The second is to compare the accuracy of the size of all the pairwise relation of the predicted ranking. Since predicted ranking results are obtained in both algorithms, the accuracy of the ranking is considered to be measured. The size of all pairwise relations in the ranking is used to represent the ranking, and it is compared with the pairwise size of the scholar's H-index 30 years later to calculate the accuracy. The calculation is shown in Formula 9.

$$\mathrm{HIT} = \frac{\delta(\mathrm{P}, \mathrm{P}')}{N * (N-1)} \qquad \text{(Formula 9)}$$

Experiment and results

This paper selects the computer science field to conduct an empirical analysis.

Data source and pre-processing

The dataset of Academic Social Networks (https://www.aminer.cn/data#Science-Knowledge-Graph) is used as literature data (Tang, Zhang, Yao, Li, Zhang& Su, 2008). For publication data, institutional data from CCF proposed Directory of CCF Recommended International Academic Conferences and Periodicals (2015, 4th edition) is used. ArnetMiner (https://www.aminer.cn/ranks/org) is used to rank the global scientific research institutions in the computer science field as basis of the index “institutional grade”. According to the ranking, if the top 1 institution is x among institutions of a scholar a_i, then its institutional reputation is y, as shown in Formula 10.

$$\mathrm{y} = \mathrm{f(x)} = \begin{cases} 1, & \text{x rank in } 001 - 200 \\ 2, & \text{x rank in } 201 - 400 \\ 3, & \text{x rank in } 401 - 600 \\ 4, & \text{x rank in } 601 - 800 \\ 5, & \text{x rank in after } 801 \\ 6, & \text{x did not enter the ranking} \end{cases} \qquad \text{(Formula 10)}$$

This paper selects the time from 1984 to 1988 (5 years) and from 1984 to 1993 (10 years) for extracting features of beginners. The extraction time for label information is set to 2014. The original label information y is H-index of beginners 30 years later, and label value of beginners (X_i, X_j) is sign(y_i- y_j), i.e. sign symbol of H-index difference between $X_i$and X_j. Pairwise ranking algorithm is pair prediction for beginners. Therefore, it is necessary to make pre-processing for eigenvalues of beginners. For beginners X_i and X_j, the input forming learning model is <X_i, X_j,sign(y_i- y_j)>. Here, a new group of features are obtained through subtracting eigenvalues of X_i and X_j, i.e. the new feature is X_{ij} and input of learning model is <X_{ij}, sign(y_i- y_j)>. $X_{ij} = (X_i^1 - X_j^1, X_i^2 - X_j^2, \ldots, X_i^m - X_j^m)$, where m is the number of features.

Empirical workflow

This paper uses two sets of feature data, i.e. those of 5 years and those of 10 years, to verify the effectiveness. These questions are to be answered: (1) Which machine learning algorithm is used? (2) Can 10-years data and 5-years data effectively identify rising stars (over 90%)? (3) Which features are of utmost importance for the identification? (4) How does the setting of different population proportions (k%) of rising stars affect the accuracy of the prediction model? (5) Can a few important features (even just H-index) be used to effectively solve the problem?

Experimental results

According to the requirements of machine learning algorithm in this study, i.e., the classification algorithm with supervised learning, we choose discriminant models, SVM (support vector machine) and XGboost, and generative model NBC (naive Bayesian). The results are shown in Table 2.

Table 2. XGboost, SVM and NBC algorithm performance measurement（10 years）

Model	*Accuracy*	*Precision*	*Recall*	*F1*	*Hit Ratio of Top K*			
					10%	*20%*	*30%*	*40%*
XGboost	0.9111	0.9112	0.9112	0.9111	0.78	0.89	0.76	0.91
SVM	0.6315	0.7039	0.6620	0.6202	0.33	0.65	0.77	0.83
NBC	0.7068	0.7377	0.7254	0.7054	0.66	0.83	0.88	0.83

The results show that the performance of XGboost is the best. XGboost is a kind of decision tree algorithm, which can score the importance of features. Table 3 shows the maximum performance of RStarRank training after adjusting main hyper-parameters.

Table 3. Performance measurement of RStarRank（5 years data/10 years data）

Model	*Accuracy*	*Precision*	*Recall*	*F1*	*Hit Ratio of Top K*			
					10%	*20%*	*30%*	*40%*
5-years data	0.8402	0.8404	0.8401	0.8401	0.70	0.77	0.73	0.89
10-years data	0.9111	0.9112	0.9112	0.9111	0.78	0.89	0.76	0.91

In terms of the performance indexes of the two models, training effect of the 10-year model is better than the 5-year model. The prediction accuracy can reach over 90% for the 10-year model. Therefore, this model can effectively predict rising stars. When ranking beginners, this model indicates a high HR for top 20% and 40% beginners, reaching over 90%. However, the prediction accuracy is 84% for the 5-year model and therefore it cannot perform effective prediction. Nevertheless, when ranking top 40% beginners, this model demonstrates an accuracy rate of about 90% HR.

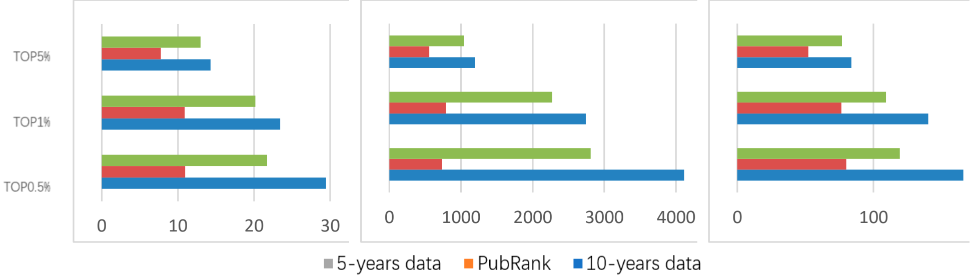
Verification for effectiveness

Method verification. There are a total of 3993 beginners in 1983. According to test data, the accuracy of the model remains stable for different years. The accuracy reaches over 90% for 10-year model and over 80% for 5-year one, and other performance indexes are similar to those obtained in 1984. So, this model indicates certain prediction effects for different years. In terms of the number of papers and citations, Top 20 scholars are leading. The mean of H-index is 27.9 and minimum H-index is 15 (as shown in Appendix 2, Table 4 shows the Top 10). Therefore, they are influential scholars. In addition, in the field of computer science, most of these scholars are professors from top universities, with multiple awards and fellowships of IEEE and ACM. Therefore, it is considered that those beginners in 1983 have grown into outstanding talents after 30-year growth, complying with standards for recognizing rising stars.

Table 4. Results of rising stars identification Top 10（year 1983）

Ranking	*Name*	*Total No. of Paper*	*Total No. of Citations Received*	*H-index*	*Academic Achievements*
1	Rakesh Agrawal	172	15757	43	National Academy of Engineering, ACM and IEEE Fellowship; Former Microsoft Technical Researcher; Former IBM Researcher
2	Brad A. Myers	272	4926	36	Professor and Senior Research Scientist at Carnegie Mellon University; 2017 ACM SIGCHI Lifetime Research Achievement Award winner; co-chaired several journals, conferences, and projects
3	Victor Vianu	126	3033	26	Professor of Computer Science, University of California, San Diego; International President of INRIA
4	Chris Faloutsos	347	9386	50	Professor of Carnegie Mellon University; Chairman/Co-Chairman of ACM SIGMOD, KDD, LDB, etc; In 2004, one of his research was rated as one of the top 50 educators in the field of information technology
5	Mark Hill	139	3894	36	Emeritus Professor of University of Wisconsin-Madison; Eckert Mozili Award (one of the most influential awards in the field of computer science) winner
6	R Tamassia	222	2471	22	Professor of School of Computer Science, Brown University; AAAS, ACM and IEEE Fellowship; IEEE Computer Society Technical Achievement Award winner
7	M. Horowitz	128	2688	27	Professor of Stanford University; IEEE and ACM Fellowship, IEEE Donald O. Pederson Award winner
8	Keith D. Cooper	76	1540	26	Professor of Rice University
9	Robert E. Kraut	135	2519	29	Professor emeritus at Carnegie Mellon University; Former professor at the University of Pennsylvania, Cornell University, and research scientist at Bell Laboratories
10	William Buxton	142	3971	36	Professor of University of Toronto, Canada; CHCCS/SCDHM Achievement Award winner

Method comparison. In order to verify advantages of proposed method, this paper compares its prediction results with existing results obtained with PubRank method. Moreover, this paper uses the data of a total of 3993 beginners in 1983 as test data for verification of the 5-year model, 10-year model and PubRank respectively.



(a) H index (b) Counts of citation paper (c) Counts of paper

Figure 2. Comparison of RStarRank and PubRank prediction results

－Comparison of prediction results in H-index, number of papers and citations. Figure 2 shows the results of Top 5%, Top 1% and Top 0.5% scholars with each method respectively. The two algorithms here rank scholars with high H-index, citations and total number of papers at top positions in prediction. Specifically, the 10-year model is 268.95%, 214.95% and 184.25% higher than PubRank in terms of H-index for Top 0.5%, Top1% and Top5% scholars. Similarly, the 10-year model is 559.04%, 347.88% and 213.85% higher than PubRank in citations; 207.69%, 183.32% and 160.56% higher than PubRank in total number of papers.

－Comparison of accuracy of all relationship pairs in predicted ranking sequence. There are 7970028 relationship pairs, since there are total 3993 beginners in 1983.

Table 5. Accuracy Comparison of All Relationships for Prediction Result Series

Model/Algorithm	*Total No. of relationship pairs*	*Correct No. of relationship pairs*	*Accuracy*
10-years Data	7970028	6816064	85.52%
PubRank	7970028	3843763	48.23%
5-years Data	7970028	6466687	81.14%

The results indicate accuracy of the proposed method reaches over 80%, much higher than 48% of PubRank. Therefore, the proposed method has certain advantages.

Feature analysis

XGboost algorithm can determine importance of each feature. The results are show in Appendix 3. For the 5-year model, “publication” and “time increment” are not considered significant. However, academic level of the co-author is more important, instead of number of co-authors. For the 10-year model, the index “institution” in 10-year model is not as important as the one in 5-year model. However, academic level of co-author is more important than number of co-authors. “Publication” and “time increment” are not completely reflected in two models, while “institution” is more important in early recognition for beginners.

Optimization

H-index as the only feature indicator. As seen from the feature analysis, the H-index is more important than other features. Can the results be analyzed based on the characteristics? The accuracy is 0.72 when H-index in the 5th year is used as the unique feature in the 5-year model, and the accuracy is 0.82 when H-index in the 10th year is used as the unique feature in the 10-year model. Compared to original model, accuracy of these two models have greatly declined.

Five feature indicators. This paper selects the 5 most important indicators to optimize the method and reduce the difficulty in data collection and the amount of calculation. The accuracy of the two data sets decreased slightly, and the hit rate of Top K ranking is basically the same, with ups and downs in some of the K rankings. In general, the prediction level of the method with only five characteristic indicators does not decrease significantly.

Table 6. Performance Measurement of RStarRank（5 features）

Model	*Accuracy*	*Precision*	*Recall*	*F1*	*Hit Ratio of Top K*			
					10%	*20%*	*30%*	*40%*
5-years data	0.8364	0.8369	0.8366	0.8364	0.69	0.79	073	0.89
10-years data	0.9064	0.9068	0.9066	0.9064	0.80	0.88	0.77	0.90

Model analysis

Feasibility analysis. RStarRank method can be used in multiple scenarios. The practicability of this method is different according to the two variables of research field and time.

Table 7. Scenario application of RStarRank

Time period *Research field*	*Same time period*	*Different (Approximate) time period*	*Different (Non-approximate) time period*
Same	Applicable	Applicable	Not applicable
Different	Not applicable	Not applicable	Not applicable

The model is trained based on original data of the computer science field and has verified the effective in computer science. Pairwise ranking algorithm is the foundation of the model, it translate the data to relative values among beginners and will not influence its feasibility due to differences in different fields, for example, the value of “citations” in biological field is higher than the one of computer science field.

Advantage analysis. The advantages mainly include: (1) The prediction accuracy reaches over 90% for the 10-year model and over 84% for the 5-year model, showing certain advantages compared to PubRank method. (2) The model can predict for different tasks, based on data conditions and using a few feature indexes. Besides, the data of these feature indexes have high availability. (3) The model provides ranking of probability of becoming rising stars. It is to compare the relative correlation of beginners in the same field, and there is no limitation for data scale. Therefore, this model can be used practically in selection of youth reviewers, recruitment of research institutions, and introduction of youth talents, etc. (4) The dependency on field shows the field feasibility of the model. Therefore, it can be used to evaluate beginners of the same field.

*Weakness analysis.*The weaknesses mainly include: (1) 5-year model depends on institutional reputation of beginners to some extent. Hence, a relatively persuasive ranking of institutions is required for different fields. (2) The accuracy of 5-year model has not reached 90%. (3)The model cannot conduct comparable ranking for beginners in different fields.

Conclusions

This paper has provided a new method for recognizing rising stars based on machine learning theory and conducted an experiment in computer science field. According to experimental results, the proposed model provides a higher prediction accuracy and can make predictions for different tasks based on data conditions or use a few feature indexes to make a prediction;

moreover, the data of these feature indexes are highly available. Without strong dependency on field, the model can be used in multiple problem-solving tasks. However, this study has some limitations. First, the achievement still emphasizes academic paper which will be investigated for general field of study. Therefore, academic paper is featured with generalization. However, some fields of study such as engineering, humanities and social sciences may investigate books and software works by a scholar. In addition, investigating personalized awards is also an important content. This paper has demonstrated the feasibility of recognizing rising stars with above problem definition from the view of overall framework, which is significant for solving general problems. However, the problem model shall be adjusted to solve some special problems. Second, it is necessary to prepare a large-scale data for early model training. The model here uses machine learning method, so that data scale is one of foundations for reaching model accuracy. Under the same model, the more training data it uses, the higher the model precision will be. Therefore, a large amount of data is required for early model training.

References

Billah, S. M. (2013). *Identifying emerging researchers using social network analysis*. Dissertations & Theses - Gradworks.

Daud, A., Ahmad, M., Malik, M. S., & Che, D. (2015). Using machine learning techniques for rising star prediction in co-author network. *Scientometrics*, 102(2), 1687–1711.

Daud, A., Aljohani, N. R., Abbasi, R. A., Rafique, Z., Amjad, T., Dawood, H., & Alyoubi, K. H. (2017). *Finding Rising Stars in Co-Author Networks via Weighted Mutual Influence*. In WWW '17 Companion Proceedings of the 26th International Conference on World Wide Web Companion (pp. 33–41).

Li, X.-L., Foo, C. S., Tew, K. L., & Ng, S.-K. (2009). *Searching for Rising Stars in Bibliography Networks*. In DASFAA '09 Proceedings of the 14th International Conference on Database Systems for Advanced Applications (pp. 288–292).

Meho L I. (2007). The Rise and Rise of Citation Analysis[J]. *Physics World*, 20(1): 32-36.

Ning, Z., Liu, Y., & Kong, X. (2017). *Social gene — A new method to find rising stars*. In 2017 International Symposium on Networks, Computers and Communications (ISNCC) (pp. 1–6).

Ning, Z., Liu, Y., Zhang, J., & Wang, X. (2017). Rising Star Forecasting Based on Social Network Analysis. *IEEE Access*, 5, 24229–24238.

Panagopoulos, G., Tsatsaronis, G., & Varlamis, I. (2017). Detecting rising stars in dynamic collaborative networks. *Journal of Informetrics*, 11(1), 198–222.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). *ArnetMiner: extraction and mining of academic social networks*. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 990–998).

Tsatsaronis, G., Varlamis, I., Torge, S., Reimann, M., Nørvåg, K., Schroeder, M., & Zschunke, M. (2011). *How to become a group leader? or modeling author types based on graph mining*. In TPDL'11 Proceedings of the 15th international conference on Theory and practice of digital libraries: research and advanced technology for digital libraries (pp. 15–26).

Wijegunawardana, P., Mehrotra, K., & Mohan, C. (2016). *Finding Rising Stars in Heterogeneous Social Networks*. In 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 614–618).

Zhang, C., Liu, C., Yu, L., Zhang, Z.-K., & Zhou, T. (2017). *Identifying the Academic Rising Stars via Pairwise Citation Increment Ranking*. In Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data (pp. 475–483).

Zhang, J., Ning, Z., Bai, X., Wang, W., Yu, S., & Xia, F. (2016). *Who are the Rising Stars in Academia*. In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (pp. 211–212).

Zhang, J., Xia, F., Wang, W., Bai, X., Yu, S., Bekele, T. M., & Peng, Z. (2016). *CocaRank: A Collaboration Caliber-based Method for Finding Academic Rising Stars*. In WWW '16 Companion Proceedings of the 25th International Conference Companion on World Wide Web (pp. 395–400).

Modeling Knowledge Diffusion in the Disciplinary Citation Network based on Differential Dynamics

Zenghui Yue[1], Haiyun Xu[2] and Guoting Yuan[3]

[1] *yzh66123@126.com*
School of Medical Information Engineering, Jining Medical University, Rizhao (China)

[2] *xuhaiyunnemo@gmail.com*
Business School, Shandong University of Technology, Zibo (China)

[3] *guotingyuan@hotmail.com*
School of Foreign Languages, Jining Medical University, Rizhao (China)

Abstract

Knowledge diffusion based on disciplinary citation is similar to disease propagation via the actual contact. Inspired by the disease-spreading model in complex networks, the paper classifies disciplines in the process of knowledge diffusion in the disciplinary citation network into 5 states, namely potential knowledge recipients, potential knowledge diffusion disciplines, knowledge diffusion disciplines, knowledge skeptics and knowledge immunes. The classifications of disciplines can transform from one state to another with certain rate (α, β, ω, γ, θ and μ respectively). Then the evolution rules of knowledge diffusion in the disciplinary citation network are made. Furthermore, the knowledge diffusion model of differential dynamics in the disciplinary citation of non-uniformity network is formed, and then the evolution process of knowledge diffusion is further discussed, so as to reveal the dynamic mechanism of knowledge diffusion in the disciplinary citation network. The research has shown that: the knowledge diffusion state evolution of disciplines in the disciplinary citation network is affected not only by the evolution states of adjacent disciplines in knowledge diffusion, but also by the relative citation weight (knowledge contact intensity).

Introduction

Knowledge diffusion, the intermediate link and intermediary process of knowledge production and knowledge application, is the exchange and dissemination of knowledge through different media, the transition of knowledge from production behavior to consumption behavior, and the transfer of knowledge from creation subject to learning subject. The ultimate goal of knowledge diffusion is the utilization and innovation of knowledge. Knowledge acquisition, distribution, transformation, innovation and application also need be realized through knowledge diffusion. In the age of Big Science, the disciplinary boundaries are becoming fuzzier with more and more frequent knowledge flow and exchange among disciplines, which greatly promote the collaboration, integration, innovation and development of the disciplines. The study of disciplinary knowledge diffusion law and its dynamic evolution mechanism is an effective way to clarify discipline development and knowledge exchange mode, predict innovation field and developing trend, and promote discipline construction and sci-tech progress.

The study of dynamic evolution mechanism of disciplinary knowledge diffusion shall be started from model building. The common idea is to learn from mature models in disciplines such as infectious diseases, complexity science, sociology, etc. The process of disciplinary knowledge diffusion is accompanied by the generation and evolution of the knowledge diffusion network, and construction of the model relies on the tangible or intangible carrier network of disciplinary knowledge diffusion. Literature is the carrier of knowledge, and the literature citation process is accompanied by the diffusion, transfer, inheritance and innovation of knowledge. Disciplinary citation network is one of the commonly used media to measure the diffusion of disciplinary knowledge. The current knowledge diffusion models include citation path model (Lu & Liu, 2013; Yu, Lu & Liu, 2014; Yan, 2014; Liu & Kuan, 2016), epidemic-like model (Bettencourt, Cinron-Arias & Kaiser, 2006; Bettencourt, Kaiser & Kaur, 2008; Kiss, Broom & Craze, 2009; Yang, Hu & Liu, 2015; Wang, Guo & Yang, 2015; Li, Zhang & Man, 2017; Yue, Xu & Yuan, 2019), network structure model (Cowan & Jonard, 2004; Kim & Park, 2009; Ozel, 2010; Liu, Rousseau & Guns, 2013; Liu, Jiang & Chen, 2015; Luo, Du & Liu, 2015), individual behavior model (Klarl, 2014), citation temporal series network model (Gao & Guan, 2012), co-citation clustering model (Wang, Zhao & Liu, 2013), etc.

Knowledge diffusion based on disciplinary citation is similar to disease propagation. In the process of disciplinary citation, knowledge exchange and diffusion can take place with literature citation among disciplines, while disease is usually propagated among organisms through physical contact, such as air, contact, matrix, blood, and so forth (Anderson & May, 1991).

In consideration of the similarity between knowledge diffusion based on literature citation and disease propagation via actual contact, and inspired by the disease-spreading model in complex networks, the paper defines 5 states of disciplines in the process of knowledge diffusion in the disciplinary citation network which can transform from one to another with a certain rate. Then the evolution rules of knowledge diffusion in the disciplinary citation network are made. Furthermore, the disciplinary knowledge diffusion model of differential dynamics in non-uniformity networks is formed and verified, and then the evolution of knowledge diffusion in the disciplinary citation network is further discussed. In this way, the effect of disciplinary citation attribute and the evolution rules of knowledge diffusion are further clarified and knowledge diffusion laws based on literature citation explored so as to reveal the dynamics mechanism of knowledge diffusion in the disciplinary citation network.

Theoretical Model

Knowledge diffusion network based on disciplinary citation

Bibliographies have long been the canonical form of data storage for recording information of scientific activities in a field or across fields. Bibliographic records contain essential summary information of the knowledge diffusion activities in a discipline, revealing 3 different yet interrelated sets of relations: (1) interactive relations among disciplines; (2) semantic relations among subject words which reveal discipline contents; (3) cognitive relations which reveal cognitive structures of disciplines.

In the process of knowledge diffusion activities based on literature citation, the sets of relations correspond to 3 different yet interrelated networks, i.e., disciplinary citation network, knowledge network and knowledge diffusion network based on disciplinary citation. The 3 networks can be described as follows:

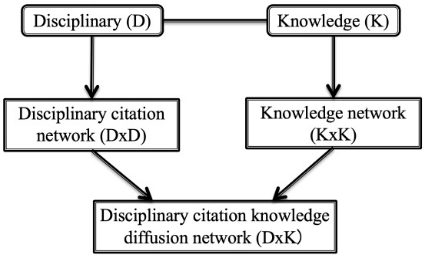

Disciplinary (D)
Knowledge (K)
Disciplinary citation network (DxD)
Knowledge network (KxK)
Disciplinary citation knowledge diffusion network (DxK)

Figure 1. Construction of knowledge diffusion network based on disciplinary citation.

Disciplinary citation network (DxD) is a network formed by mutual citation of literature among disciplines. Its unit value represents the frequency of mutual citation of those disciplines and reflect the integration degree of disciplinary knowledge.

Knowledge network (KxK) is a network formed by the co-occurrence among knowledge. Its unit value represents the number of times that different knowledge appears in the same literature in pairs, reflecting the correlation relationship of different knowledge units.

Disciplinary citation knowledge diffusion network (DxK) is formed by the co-occurrence of disciplines and knowledge. The unit value represents the amount of corresponding knowledge in a certain discipline and reflects the distribution or diffusion of knowledge in the discipline (i.e., the cognitive state of the discipline). Construction of the above networks is the basis for analyzing and exploring the essence of knowledge diffusion in the disciplinary correlation network with the aid of literature citation.

Knowledge diffusion process of the disciplinary citation network

Many studies of the behavior dynamics mechanism of propagation and diffusion are based on the classical SIR epidemic model, in which individual diffusion states are defined to be the susceptible state (S), infective state (I), and recovered/immune state (R). They can transform from one state to another in a unidirectional way with a certain rate. During the actual evolution process of knowledge diffusion in the disciplinary citation network, however, there are 3 special phenomena: (1) When receiving certain knowledge, a discipline tends not to spread the knowledge immediately. Instead, there exists a certain period of time in which it digests the knowledge. After the discipline absorbs and converts the knowledge into intrinsic knowledge, it will spread the knowledge to others. (2) The citation behavior is often affected by a variety of factors, such as the rational factors, social factors and random factors, etc. (Ma & Wu, 2009). There are passive or negative citations in some disciplinary literature, where the opinions or works in those citations are criticized because of certain shortcomings or even negated. Then the knowledge in the discipline might be questioned and spread the knowledge further hindered. (3) As the evolution and development of science, a discipline may not diffuse knowledge anymore. However, this kind of immunity is temporary for some disciplines, because they will accept and diffuse the knowledge again in the future. Therefore, this paper argues that the SEIZRS epidemic model, which has a certain latent state (E), skeptical state (Z), and feedback mechanism (R to S), is more consistent with the nature of the knowledge diffusion process in the disciplinary citation network, and is more suitable to function as the basic model of differential dynamics for knowledge diffusion in the disciplinary citation network.

States of disciplines in the knowledge diffusion process of the disciplinary citation network

In terms of characteristics of disciplinary citation and diversity of citation motives, we define that disciplines in different states of knowledge diffusion in the disciplinary citation network fall into 5 categories:

(1) Potential knowledge recipients (S): disciplines which have not known the knowledge or have known but not acquired it yet.
(2) Potential knowledge diffusion disciplines (E): disciplines which have acquired the knowledge but not diffused it yet.
(3) Knowledge diffusion disciplines (I): disciplines which have mastered the knowledge and are diffusing it to the potential knowledge recipients.
(4) Knowledge skeptics (Z): disciplines which have acquired the knowledge but disapprove of it and even question it.
(5) Knowledge immunes (R): disciplines which have acquired the knowledge but are immune to it now. They have lost interest in the knowledge and will not continue to diffuse it.

Process of knowledge diffusion in the disciplinary citation network

In the initial phase, there are only small numbers of knowledge diffusion disciplines in the network, while others are potential knowledge recipients. The number of potential knowledge diffusion disciplines, knowledge skeptics, and knowledge immunes is zero. As time goes on, knowledge begins to be diffused in the following ways:

When knowledge diffusion disciplines transmit knowledge to potential knowledge recipients, the recipients begin to accept it with a certain rate (α) and become potential knowledge diffusion disciplines, while some potential knowledge recipients who disapprove of the knowledge become knowledge skeptics with θ. If potential knowledge diffusion disciplines are interested in the knowledge, they will continue to diffuse it with β and become new knowledge diffusion disciplines. Knowledge diffusion disciplines lose interest in the knowledge with ω and turn into knowledge immunes. Meanwhile, when knowledge skeptics thoroughly reject the knowledge with μ, they also transform into knowledge immunes. Knowledge immunes become knowledge recipients again with γ and take in the knowledge to which they have been immune.

Evolution rules of knowledge diffusion in the disciplinary citation network

At a certain time in the knowledge diffusion process in the disciplinary citation network, which could be marked as t, a discipline can only be in one of the states mentioned above. At time t, we can define the proportion of disciplines that are in a certain knowledge diffusion state as follows.

(1) s(t): the proportion of potential knowledge recipients to all disciplines in different states of knowledge diffusion at time t.
(2) e(t): the proportion of potential knowledge diffusion disciplines to all disciplines in different states of knowledge diffusion at time t.
(3) i(t): the proportion of knowledge diffusion disciplines to all disciplines in different states of knowledge diffusion at time t.
(4) z(t): the proportion of knowledge skeptics to all disciplines in different states of knowledge diffusion at time t.
(5) r(t): the proportion of knowledge immunes to all disciplines in different states of knowledge diffusion at time t.

Here, s(t)+e(t)+i(t)+z(t)+r(t)=1.

According to the above description of the knowledge diffusion process in the disciplinary citation network, a schematic paradigm of the dynamic state evolution rules of disciplines in knowledge diffusion is presented as follows (Figure 2).

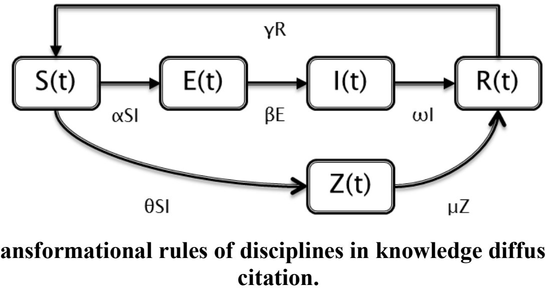


Figure 2. State transformational rules of disciplines in knowledge diffusion of disciplinary citation.

When there is citation relationship between potential knowledge recipients and knowledge diffusion disciplines in the process of disciplinary citation, the former could become potential knowledge diffusion disciplines with the rate of α, and become knowledge skeptics with θ; potential knowledge diffusion disciplines become knowledge diffusion disciplines with β; knowledge diffusion disciplines turn into knowledge immunes with ω; knowledge skeptics turn to be knowledge immunes with μ; and knowledge immunes generate feedback with γ, and become potential knowledge recipients.

Modeling of knowledge diffusion in the disciplinary citation network

The disciplinary citation network is a kind of complex networks with obvious scale-free features. Disciplines with different citation weight (knowledge contact intensity) play different roles in the process of knowledge diffusion. According to the evolution rule of knowledge diffusion in the disciplinary citation network mentioned above, the disciplinary citation knowledge diffusion model of differential dynamics in the non-uniformity network is formed on the basis of the disease-spreading equation of SEIZRS and structural attribute of the citation network. The SEIZRS model equations are as follows.

$$\begin{cases} \frac{ds_i(t)}{dt} = -\alpha\sum_j w_{ji} i_j(t) s_i(t) - \theta\sum_j w_{ji} i_j(t) s_i(t) + \gamma r_i(t) \\ \frac{de_i(t)}{dt} = \alpha\sum_j w_{ji} i_j(t) s_i(t) - \beta e_i(t) \\ \frac{di_i(t)}{dt} = \beta e_i(t) - \omega i_i(t) \\ \frac{dz_i(t)}{dt} = \theta\sum_j w_{ji} i_j(t) s_i(t) - \mu z_i(t) \\ \frac{dr_i(t)}{dt} = \omega i_i(t) + \mu z_i(t) - \gamma r_i(t) \end{cases} \qquad (1)$$

In the equation set above, t is the time step; $s_i(t)$, $e_i(t)$, $i_i(t)$, $z_i(t)$, and $r_i(t)$ indicate the proportion of disciplines at time t, and their knowledge diffusion states are S, E, I, Z, and R, respectively; $w_{ji} = \frac{W_{ji}}{\sum_j W_{ji}}$ represents the standardized weight of knowledge diffusion from discipline "j" to "i", and W_{ji} is the number of citations from discipline "i" to "j".

Model simulation and verification

Based on Model 1, the simulation experiment of the disciplines' state evolution of knowledge diffusion in the disciplinary citation network from t_1 to t_v is conducted by employing MATLAB. The initial conditions of the model are that when t=0, i(0)>0, s(0)=1-i(0)>0, e(0)=0, z(0)=0, r(0)=0.

Parameter estimation

We adjust the parameters (α, β, ω, γ, θ and μ) constantly to gain the best fitting model, namely, the supreme fitness between the output of the model and the actual value that is the proportion of knowledge diffusion disciplines in the disciplinary citation network. According to the distributing characteristics of the theoretical and practical data values, the Likelihood function is employed in this paper for parameter estimation.

The Likelihood function in this model is:

$$L = \frac{1}{v}\sum_{t=t_1,t_2,\ldots\ldots,t_v}\left(-\log\left(\prod_{j=1}^{N}\left(i_j(t)\right)^{Y_j(t)}\left(1-i_j(t)\right)^{1-Y_j(t)}\right)\right)$$

（2）

where v is the total number of time steps. $Y_j(t)$ is an indicator function with a value of one, denoting that a discipline j is spreading knowledge at time t, and a value of zero denoting that j is not a knowledge diffusion discipline. According to Formula 2, the set of parameters that gain the minimum value for L is the optimal parameter set of the fitting model.

Model building

The optimal parameters of the fitting model will be plugged into Model 1 to form the knowledge diffusion model of differential dynamics eventually.

Model verification

The optimal fitting model can continue to be used through iteration of time steps to predict the changing trend of the quantitative proportion of the knowledge diffusion disciplines. By comparing the gained data with the actual data from t_{v+1} to t_m, the consistency can be further checked. If they are consistent, the model is correct; otherwise, it will be revised.

Empirical Research

Selection of research object and data acquisition

As a new sociological research paradigm, social network theory has been continuously deepened in different disciplines such as Psychology, Sociology, Anthropology and so on, and knowledge transmission and diffusion in this field are very active (Wang & Liu, 2007). Social media is a platform for users' social communication and information sharing. It has become an important kind of media for obtaining current information, interpersonal communication, self-expression, social sharing and social participation, through which people form certain loose or close social network connections (Xu & Jin, 2013). There are many similarities between social network and social media in theory, technical methods and applications, so the studies of them present an overlapping, mutually permeating, and inseparable trend.

Therefore, the knowledge point "social media" in the field of social network is selected as the research object to form and verify the knowledge diffusion model of differential dynamics that simulates the knowledge diffusing process of social media in the disciplinary citation network of social network, and the research results are further analyzed and explained. Because the knowledge point "social media" in the field of social network emerged in 2008, the research period of this model is set from 2008 to 2019.

In this paper, TS="social network*" is set as the searching strategy to collect articles in SCI-EXPANDED and SSCI of Web of Science from 2008 to 2019. After data cleaning, 35,397 articles are obtained. In terms of classification of disciplines according to Web of Science Categories and on the basis of the citation data, the multimap of journals and disciplines is used and temporal evolution networks of knowledge diffusion in disciplinary citation during the period is extracted to construct the knowledge diffusion network based on disciplinary citation. The classification of disciplines is not exclusive. In the multimap of journals and disciplines, when a journal has mapping relationships with more than one disciplines, the journal is classified into all those disciplines.

State evolution of knowledge diffusion disciplines

In this paper, we define disciplines in the state of knowledge diffusion disciplines at a certain time as the disciplines that publish papers containing the knowledge point "social media" in the social network field at that time. The total number of members in the disciplinary citation network is 226. The quantity and proportion of knowledge diffusion disciplines from 2008 to 2019 are shown in Table 1.

Table 1. Quantity and proportion of knowledge diffusion disciplines during 2008–2019.

Year	Quantity	Proportion
2008	3	0.013274
2009	13	0.057522
2010	33	0.146018
2011	41	0.181416
2012	66	0.292035
2013	77	0.340708
2014	93	0.411504
2015	104	0.460177
2016	92	0.40708
2017	102	0.451327
2018	116	0.513274
2019	118	0.522124

It can be seen from the table that the proportion of disciplines in the state of knowledge diffusion disciplines (in other words, disciplines that are diffusing knowledge) is growing continually (from 1.3274% to 52.2124%), though the proportion declines in 2016.

Model construction and validation

In the original state (i.e., in 2008), there were 3 knowledge diffusion disciplines of social media in the network (i.e., Computer Science, Artificial Intelligence; Computer Science, Information Systems; and Computer Science, Theory & Methods). The others were all potential knowledge recipients, without any potential knowledge diffusion disciplines, knowledge skeptics and knowledge immunes.

Parameter determination

On the basis of Model 1, we simulate the theoretical model by using MATLAB, adjust parameters (α, β, ω, γ, θ and μ) with a step length of 0.1 and values from 0 to 10, and calculate Maximum-Likelihood estimation (L) between the output values of the model with different parameters and actual values (i.e., proportions of knowledge diffusion disciplines of social media) from 2008 to 2017, respectively, to determine the optimal parameters of the fitting model (Table 2).

Table 2. Optimal parameter set of the fitting model and its L value.

α	β	ω	γ	θ	μ	L
2	10	0.2	0.9	1	0.4	6.51677095

Model building

Evolution curves of the theoretical and actual proportions of knowledge diffusion disciplines in the optimal fitting model are shown in Figure 3.

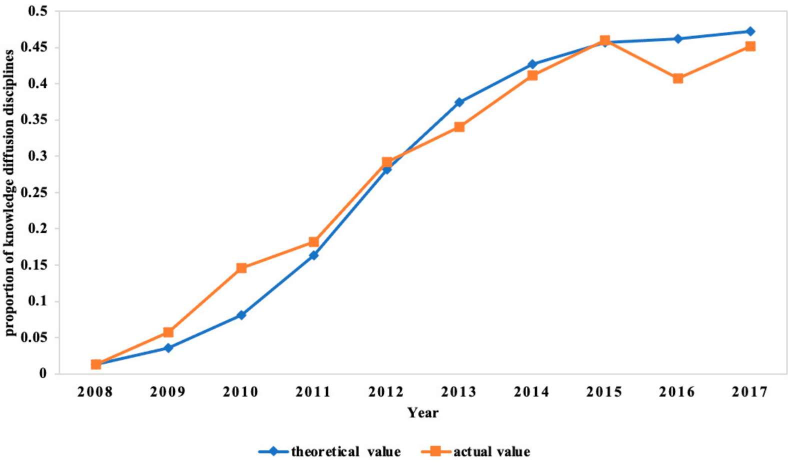


Figure 3. Evolution curves of theoretical and actual proportion of knowledge diffusion disciplines with the optimal fitting parameters.

In the disciplinary citation network of social network, under the conditions that potential knowledge recipients become potential knowledge diffusion disciplines with a state transition parameter of 2 and knowledge skeptics with that of 1, potential knowledge diffusion disciplines become knowledge diffusion disciplines with a parameter of 10, knowledge diffusion disciplines turn into knowledge immunes with a parameter of 0.2, knowledge skeptics become knowledge immunes with a parameter of 0.4, and knowledge immunes revert to being potential knowledge recipients with a parameter of 0.9, the model can simulate the knowledge diffusion process of social media in the disciplinary citation network of social network and reflect the dynamic nature and inherent law of knowledge diffusion in the disciplinary citation network. Therefore, the knowledge diffusion model simulating the knowledge diffusion of social media in the disciplinary citation network of social network can be described as follows.

$$\begin{cases} \frac{ds_i(t)}{dt} = -2\sum_j w_{ji} i_j(t) s_i(t) - \sum_j w_{ji} i_j(t) s_i(t) + 0.9 r_i(t) \\ \frac{de_i(t)}{dt} = 2\sum_j w_{ji} i_j(t) s_i(t) - 10 e_i(t) \\ \frac{di_i(t)}{dt} = 10 e_i(t) - 0.2 i_i(t) \\ \frac{dz_i(t)}{dt} = \sum_j w_{ji} i_j(t) s_i(t) - 0.4 z_i(t) \\ \frac{dr_i(t)}{dt} = 0.2 i_i(t) + 0.4 z_i(t) - 0.9 r_i(t) \end{cases} \qquad （3）$$

The model illustrates that the state evolution of knowledge diffusion disciplines in the disciplinary citation network is affected not only by the evolution states of adjacent disciplines in knowledge diffusion, but also by the relative citation weight (knowledge contact intensity).

Model verification

In accordance with Model 3, the proportions of disciplines that are diffusing knowledge from 2018 to 2019 are predicted by the iteration of time steps (Table 3).

Table 3. Predicted value of the proportion of knowledge diffusion disciplines from 2018 to 2019 in theory.

Year	Predicted value	Actual value	Deviation
2018	0.48261152	0.51327434	-5.974%
2019	0.48641531	0.52212389	-6.839%

The table shows that the deviations between the theoretical value and actual value of the proportion of knowledge diffusion disciplines predicted from 2018 to 2019 are -5.974% and -6.839%, respectively. The closeness between the values illustrates the validity of the model.

A comparison of SEIZRS model and SIR model

In order to further verify the validity of the introduced knowledge latent mechanism, skeptical mechanism and feedback mechanism in the paper, a comparison analysis is performed between the evolutionary fitting effects of SEIZRS model and classic SIR model from 2008 to 2019 (Figure 4).

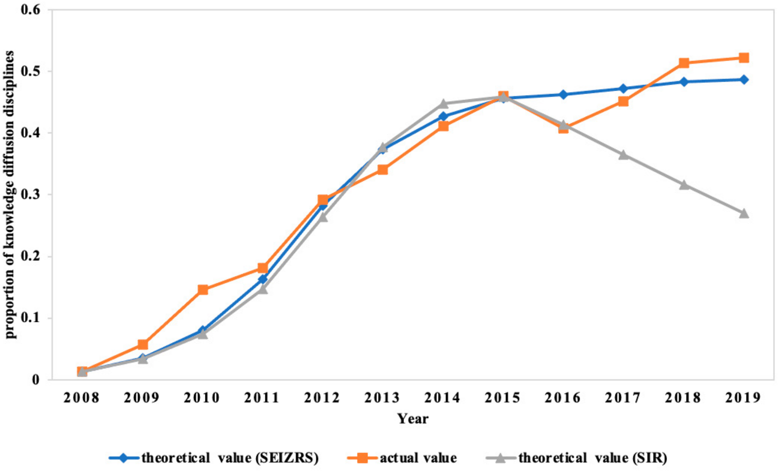


Figure 4. Evolution curves of theoretical and actual proportion of knowledge diffusion disciplines in the optimal fitting model with SEIZRS and SIR type transmission respectively.

Before 2016, the SIR model could relatively effectively reflect diffusion process of the knowledge points "social media" among disciplines by drawing on literature citation in the field of social network. However, its fitting ability has declined rapidly since 2017, and the gap between the theoretical output value of the model and the actual proportion value of knowledge diffusion disciplines has become increasing larger.

From 2008 to 2019, the performance of SEIZRS model is more consistent with the knowledge diffusion process in the disciplinary citation network, which shows that the knowledge latent mechanism, skeptical mechanism and feedback mechanism introduced in this paper can more effectively reveal the essential mechanism and dynamic mechanism of knowledge diffusion in the disciplinary citation network.

Evolution analysis of knowledge diffusion in the disciplinary citation network

From 2008 to 2019, the state evolution of disciplines diffusing the knowledge of social media in the disciplinary citation network of social network is shown in Figure 5.

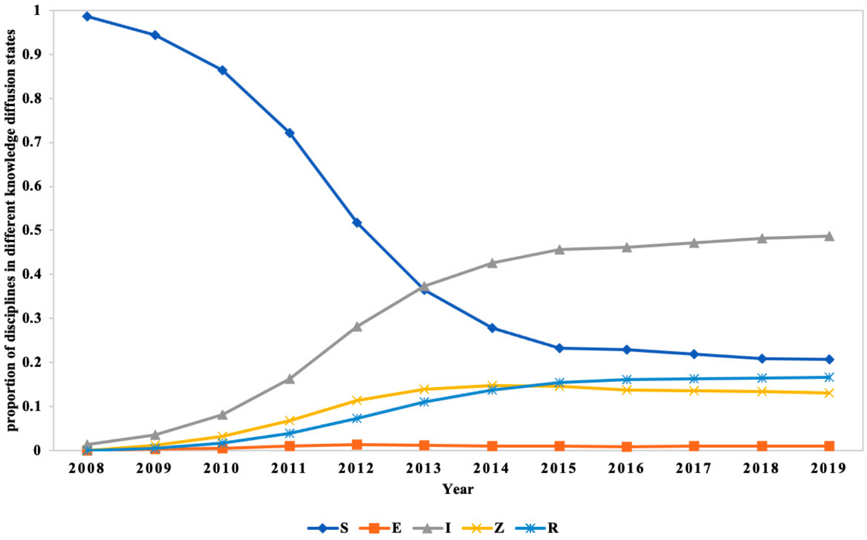


Figure 5. Evolution curves of disciplines in different knowledge diffusion states in the disciplinary citation network.

With the passage of time, it is shown from the perspective of the overall evolution trend of knowledge diffusion as follows: the proportion of potential knowledge recipients in the disciplinary citation network decreases continuously; the proportion of knowledge diffusion disciplines and knowledge immunes is increasing; the proportion of knowledge skeptics climbs first and then decreases gradually; and the proportion of potential knowledge diffusion disciplines shows a slight growth trend. From the perspective of the evolution speed, it is shown that the proportions of potential knowledge recipients, knowledge diffusion disciplines, knowledge skeptics and knowledge immunes in the disciplinary citation network all change rapidly at first and then slow down, while the change of the proportion of potential knowledge diffusion disciplines is relatively insignificant. In terms of the evolution acceleration of knowledge diffusion, the changes of potential knowledge recipients are the fastest, followed by knowledge diffusion disciplines, knowledge skeptics and knowledge immunes in succession, and changes of potential knowledge diffusion disciplines are the slowest.

In addition, according to the evolution trend of the proportion of knowledge diffusion disciplines (showing a logical growth trend), it can be judged that the diffusion of the knowledge point "social media" in the field of social network has entered a relatively mature stage, and the diffusion scale has been steady.

Discussion

From the perspective of evolution of knowledge diffusion process, differential dynamic model of knowledge diffusion in the disciplinary citation network is built in this paper to further reveal the particular dynamics mechanism of knowledge diffusion in the disciplinary citation network according to the attribute of the disciplinary citation network by improving traditional dynamics model, such as introducing knowledge latent mechanism, skeptical mechanism and feedback mechanism and considering the network heterogeneity. The research has shown that:

(1) The hypothesis that knowledge can be diffused with the aid of disciplinary citation is justified, and the knowledge diffusion model constructed in this paper is rational and reasonable.

(2) The knowledge diffusion state evolution of disciplines in the disciplinary citation network is affected not only by the evolution states of adjacent disciplines in knowledge diffusion with citation relationship, but also by the relative citation weight (knowledge contact intensity).
(3) When the state transition parameter meets the conditions that α=2, β=10, ω=0.2, γ=0.9, θ=1, and μ=0.4, the knowledge diffusion model of differential dynamics can better simulate the knowledge diffusion process of social media in the disciplinary citation network of social network.
(4) The diffusion of the knowledge point "social media" in the field of social network has entered a relatively mature stage, and the diffusion scale has been steady.
Knowledge diffusion process of "social media" in the disciplinary citation network of social network is taken as an example to verify the validity of SEIZRS model. However, there are still some questions for further exploration and discussion. What about the applicability of the model for various knowledge points of different fields? Is the speed of diffusion constant across knowledge points? What about the effect of changing parameters on the knowledge diffusion process in the disciplinary citation network?
In order to achieve different evolution effects of knowledge diffusion and control the diffusion process in the disciplinary citation network, the structure and parameters in state transition of disciplinary citation network could be regulated through various means, so as to promote or inhibit knowledge diffusion among disciplines. In addition, the future trend of knowledge diffusion could be predicted timely and accurately according to the current disciplinary citation network, so that the management and regulation of knowledge diffusion in the disciplinary citation network could be realized from the perspective of differential dynamics according to actual needs.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No. 71704063), National Natural Science Foundation of China (Grant No. 71704170), Shandong Provincial Natural Science Foundation, China (Grant No. ZR2015GL015), and Tai Shan Scholar Foundation in Shandong province (2021-2025).

References

Anderson, R.M. & May, R.M. (1991). *Infectious diseases of humans: Dynamics and control. Oxford*, UK: Oxford University Press.

Bettencourt, L.M.A., Cinron-Arias, A. & Kaiser, D.I., et al. (2006). The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models. *Physica A: Statistical Mechanics and its Applications*, 364, 513-536.

Bettencourt, L.M.A., Kaiser, D.I. & Kaur, J., et al. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics*, 75(3), 495-518.

Cowan, R. & Jonard, N. (2004). Network structure and the diffusion of knowledge. J*ournal of Economic Dynamics & Control*, 28, 1557-1575.

Gao, X. & Guan, J.C. (2012). Network model of knowledge diffusion. *Scientometrics*, 90(3), 749-762.

Klarl, T. (2014). Knowledge diffusion and knowledge transfer revisited: two sides of the medal. *Journal of Evolutionary Economics*, 24(4), 737-760.

Kim, H. & Park, Y. (2009). Structural effects of R&D collaboration network on knowledge diffusion performance. *Expert Systems with Applications*, 36(5), 8986-8992.

Kiss, S.Z., Broom, M. & Craze, P., et al. (2009). Can epidemic models describe the diffusion of topics across disciplines?. *Journal of Informetrics*, 4(1), 74-82.

Li, J., Zhang, Y. & Man, J., et al. (2017). SISL and SIRL: two knowledge dissemination models with leader nodes on cooperative learning networks. *Physica A: Statistical Mechanics and its Applications*, 468, 740-749.

Liu, J.S. & Kuan, Chung-Huei. (2016). A new approach for main path analysis: decay in knowledge diffusion. *Journal of the American Society for Information Science and Technology*, 67(2), 465-476.

Liu, X., Jiang, S. & Chen, H.C., et al. (2015). Modeling knowledge diffusion in scientific innovation networks: an institutional comparison between China and US with illustration for nanotechnology. *Scientometrics*, 105(3), 1953-1984.

Liu, Y.X., Rousseau, R. & Guns, R. (2013). A layered framework to study collaboration as a form of knowledge sharing and diffusion. *Journal of Informetrics*, 7(3), 651-664.

Lu, L.Y.Y. & Liu J.S. (2013). An innovative approach to identify the knowledge diffusion path: the case of resource-based theory. *Scientometrics*, 94(1), 225-246.

Luo, S., Du, Y. & Liu, P., et al. (2015). A study on coevolutionary dynamics of knowledge diffusion and social network structure. *Expert Systems with Applications*, 42(7), 3619-3633.

Ma, F. & Wu, Y.S. (2009). A survey study on motivations for citation. *Journal of Intelligence*, 28(6), 9-14+8.

Ozel, B. (2010). *Scientific collaboration networks: knowledge diffusion and fragmentation in Turkish management academia*. Istanbul: Istanbul Bilgi University.

Wang, J.P., Guo, Q. & Yang, G.Y., et al. (2015). Improved knowledge diffusion model based on the collaboration hypernetwork. *Physica A: Statistical Mechanics and its Applications*, 428, 250-256.

Wang, X.J. & Liu, H.L. (2007). An analysis of knowledge chain based on social network theory. *Journal of Intelligence*, 2, 18-21.

Wang, X.Z., Zhao, Y.J. & Liu, R., et al. (2013). Knowledge-transfer analysis based on co-citation clustering. *Scientometrics*, 97(3), 859-869.

Xu, Y. & Jin, J.B. (2013). The development of social media and its social impact. *Media*, 6, 10-13.

Yan, E.J. (2014). Finding knowledge paths among scientific disciplines. *Journal of the American Society for Information Science and Technology*, 65(11), 2331-2347.

Yang, G.Y., Hu, Z.L. & Liu, J.G. (2015). Knowledge diffusion in the collaboration hypernetwork. *Physica A: Statistical Mechanics and its Applications*, 419, 429-436.

Yu, X., Lu, L.Y.Y. & Liu J.S., et al. (2014). Knowledge diffusion path analysis of data quality literature: a main path analysis. *Journal of Informetrics*, 8(3), 594-605.

Yue, Z.H., Xu, H.Y. & Yuan, G.T., et al. (2019). Modeling Study of Knowledge Diffusion in Scientific Collaboration Networks based on Differential Dynamics: A Case Study in Graphene Field. *Physica A: Statistical Mechanics and its Applications*, 524, 375-391.

Network Analysis of Econometric Society Fellows

Tolga Yuret[1]

[1] *tyuret@gmail.com*
Faculty of Management, Department of Economics, Istanbul Technical University, Macka,
Istanbul, 34367 (Turkey)

Abstract

We analyze co-author and co-worker networks of 981 Econometric Society Fellows who were elected between 1933 and 2017. The vast majority of Fellows are connected either directly or indirectly. The percentage of connected Fellows has increased through time.

Introduction

Econometric Society was founded in 1933 by a group of distinguished economists. New Fellows of the Society are elected annually by the incumbent Fellows (Hamermesh & Schmidt 2003). Most top-notch economists are Econometric Society Fellows. For example, the vast majority of the Nobel Prize recipients in economics were previously elected as Fellows of the Society (Chan & Torgler 2012).

In this paper, we analyze the connections among all 981 Fellows who have been elected from 1933 to 2017. The Fellows work in a few prestigious institutions so that they have many co-workers who are also Fellows of the Society. The Fellows also publish joint papers. Using these relations, we establish co-worker and co-author networks. We find that the vast majority of Fellows are connected either directly or indirectly in both networks.

We also analyze the development of the networks through time. As the years pass, the number of Fellows has increased and the generational gap among members has widened. Despite this fact, we find that the percentage of connected Fellows has increased through the years.

Data

We collected data for this study and Yuret (2020) at the same time. The latter study analyzes co-author and co-worker networks of 3,587 researchers who published in the top five economics journals in the last ten years. 367 of these researchers are also Econometric Society Fellows. Naturally, the data descriptions of both studies are similar.

We collected publications of all Fellows to establish their co-author network. We searched the names of Fellows from Econ-lit data-base. There is room for name confusion since some authors share the same surname and first name initials. To refrain from this problem, we matched address information in publications with biographical information in CVs. We included articles, review articles, conference proceedings, notes, comments, corrections, replies, and discussions but excluded book reviews, editorials, introductory materials, and announcements.

We collected biographical information of all Fellows to establish their co-worker network. We mainly relied on Fellows' CVs in their home pages and web sites such as Prabook, Wikipedia, and LinkedIn. We only included primary academic positions since visiting information was missing in many cases. We used address information from publications when we could not get biographical information from other sources.

We could not find information about the PhD degrees of 63 Fellows many of whom may not have gotten the degree. There are 21 Fellows that we could not get the institution for their undergraduate degrees. Their biographical information either stated the country of their undergraduate degree or stated their country of birth that we used in place of the country of their undergraduate degree.

Trends in Publications and Residence

Figure 1 shows the cumulative number of Fellows through the years. We do not exclude the retired or deceased Fellows from any of our calculations. There were elections every year except for 1934, 1936, and 1940 through 1943, so the number of Fellows increased every year except for these years. The number of Fellows increased at a higher rate after 1970.

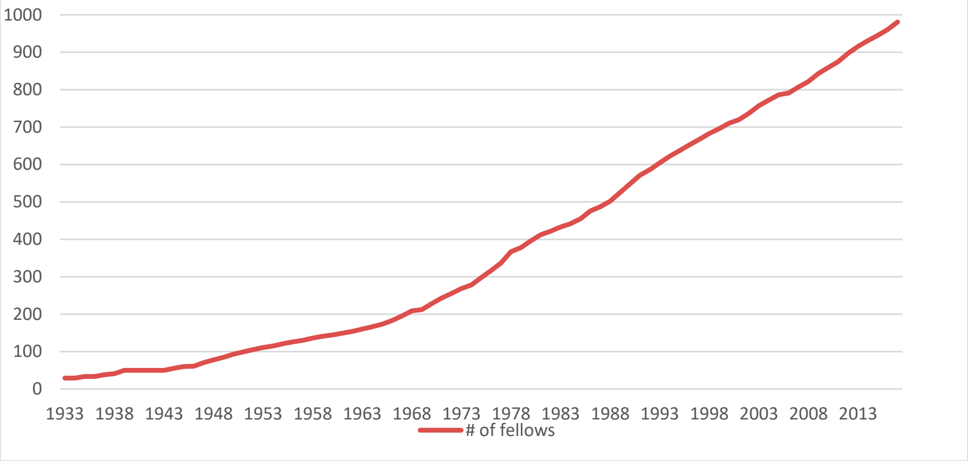


Figure 1. Cumulative number of Fellows through time

Figure 2 gives the age of the Fellows when they are elected. The highest frequency of Fellows are between ages 33 and 50. After age 50, the frequency of Fellows decreases for older age intervals. Therefore, most Fellows were elected before their retirement age.

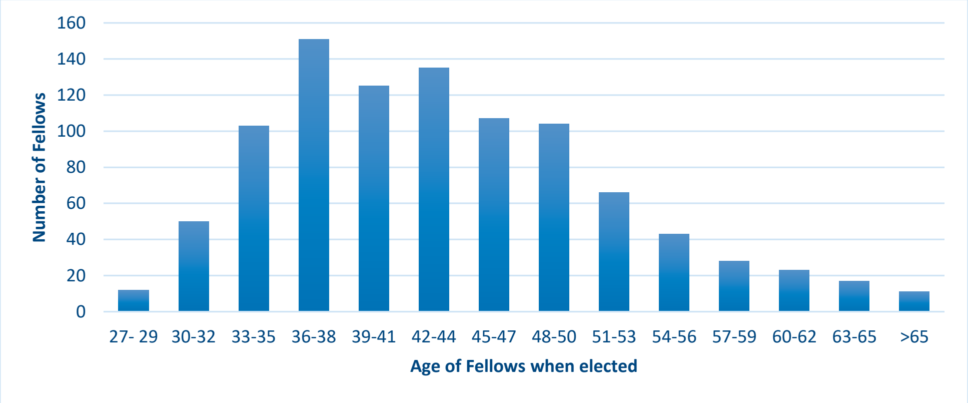


Figure 2. Age of Econometric Society Fellows when elected

Table 1 gives the number of Fellows by the country that they attained their education degrees, and the country of their workplace when elected. There are half Fellows in the workplace category because some Fellows were holding two primary positions in two different countries when they were elected. Ten countries cover more than 80% of the undergraduate and PhD degrees, and workplaces when elected. More than 40% of the undergraduate degrees and more than 70% of the PhD degrees were obtained from the United States. More than 60% of the Fellows were elected while working in a US institution.

Table 1. Number of Fellows by the country of their education and workplace when elected

Undergraduate Degree			PhD Degree			Workplace when elected		
Rank	Country	# of Fellows	Rank	Country	# of Fellows	Rank	Country	# of Fellows
1	USA	414 (42.2%)	1	USA	654 (71.2%)	1	USA	600 (61.2%)
2	UK	94	2	UK	80	2	UK	91
3	France	76	3	France	44	3	France	55
4	Israel	46	4	Netherlands	20	4	Israel	27
5	Canada	39	5	Germany	17	5	Japan	23.5
6	Italy	33	6	Israel	16	6	Canada	17
7	Japan	30	7	Sweden	13	7	Germany	17
8	India	26	8	Canada	10	8	Australia	13.5
9	Germany	22	8	Australia	8	9	Italy	13.5
10	Netherlands	20	10	Austria	8	10	Netherlands	13
Total Top 10		800			870			870.5
Total Fellows		981			918			981
Percentage Top 10		81.5			94.7			88.7

Figure 3 shows the percentage of Fellows who were educated and worked when elected in the USA through time. Percentage in a certain year is computed by considering all Fellows who have been elected up to that year. We see that all three percentages were lower in the beginning years of the Society, and there is an increasing trend up until the 1990s. The percentage of undergraduate degrees from the USA decreases, but the other two percentages are stable after the 1990s.

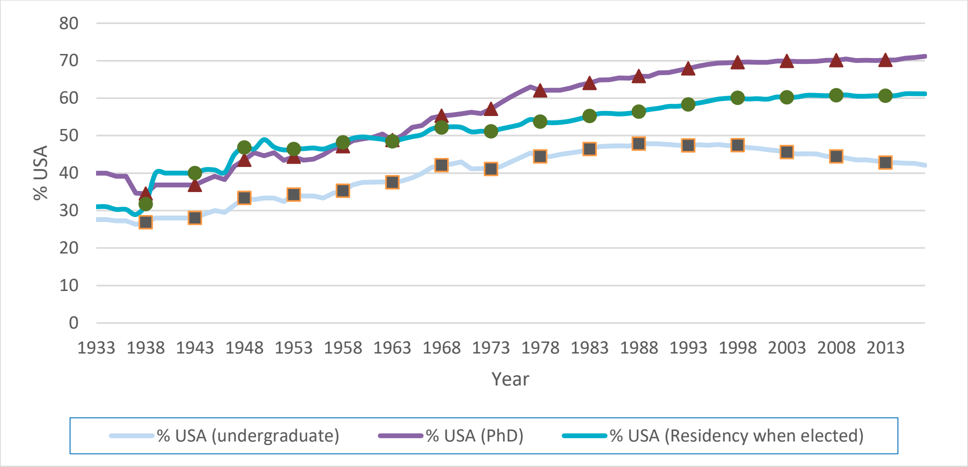


Figure 3. Percentage of Fellows who are educated and worked when elected in the USA.

There is also concentration at the institutional level as we see from Table 2. More than half of the Fellows take their PhD degrees from ten institutions, and more than one-third of the Fellows were working in just ten institutions when they were elected.

Table 2. Number of Fellows by the institution of their education and of workplace when elected

Undergraduate Alumni			PhD Alumni			Workplace when elected		
Rank	Institution	# of Fellows	Rank	Institution	# of Fellows	Rank	Institution	# of Fellows
1	Harvard U	51	1	Harvard U	98	1	MIT	44
2	Cambridge U	34	2	MIT	90	2	Stanford U	42.5
3	Hebrew U	31	3	U Chicago	62	3	U Chicago	41
4	E. Polytech.	26	4	Stanford U	60	4	Harvard U	37.5
5	Yale U	25	5	UC Berkeley	48	4	Northwestern U	37.5
6	U Chicago	21	6	Princeton U	43	6	Princeton U	34.5
7	Princeton U	20	7	Yale U	41	7	U Pennsylvania	33.5
8	ENS	19	8	U Minnesota	27	8	Yale U	33
9	UC Berkeley	18	9	LSE	26	9	LSE	29.5
10	U Tokyo	16	10	Cambridge U	22	10	UC Berkeley	25
Total Top 10		261			517			358
Total Fellows		960			918			981
Percentage Top 10		27.2			56.3			36.5

Co-author and co-worker networks

We assign each of the 981 Fellows to a node. In the co-author network, we connect two nodes with an edge if they have ever been co-authors in a paper. In the co-worker network, we connect two nodes if they worked in the same institution for three years at the same time. Because most PhD students work as teaching or research assistants during their graduate studies, we included this period when we established co-worker links.

Table 3 presents statistics of both networks as of 2017. Average degree is the average number of direct neighbors in a network. Therefore, each Fellow has on average 4.4 co-authors and 57.2 co-workers who are also Fellows of the Society. Isolated nodes are the nodes that do not have any direct connection to other nodes. Therefore, 18.3% of Fellows did not write any joint papers with any other Fellow, whereas only 3.0% of the Fellows were not co-workers with any other Fellow.

Network density divides the number of edges in a network to the maximum possible number of edges in that network. Since both networks have 981 nodes, the maximum possible number of edges in both networks is C(981,2) where C is the combinatorics function. In other words, the denominators for both networks are the same. Therefore, we can say that there are more than ten times (0.058/0.005) more edges in the co-worker network.

Giant component is the largest component where all nodes are connected either directly or indirectly. For example, two nodes are connected even if they only have a common co-author who is also a Fellow. The distance, in this case, is two since we have two edges between these two nodes. The maximum distance between two nodes in the giant component is 11 in both networks. The average distance is 4.4 in the co-author network and 2.4 in the co-worker network.

Yuret (2020) also finds that the co-worker network is much denser than the co-author network for the researchers who have published in the top five economics journals in the last ten years. Giant components are larger in that study, possibly because Fellows span more generations than the authors of the top five economics journals in the last ten years.

Table 3. Network statistics for co-author and co-worker networks

	Co-author	Co-worker
Nodes	981	981
Average Degree	4.4	57.2
Isolated Nodes	180 (18.3%)	29 (3.0%)
Network Density	0.005	0.058
Giant Component	783 (79.8%)	939 (95.7%)
Distance in Giant Component		
Maximum	11	11
Average	4.4	2.4

Figure 4 gives the size of the giant component and the size of isolated nodes as a percentage of all nodes through time. We start with 1937 and repeat the exercise for every decade. For example, when we compute the network for 1957, we consider Fellows who have been elected up until 1957 and use their co-authorship and co-worker relations that have been established up until 1957. We see that the size of giant component has grown and the size of isolated nodes has become smaller relative to the number of nodes through time.
Goyal et al. (2006) compute the giant component for co-authors in all journals indexed in Econlit. They also find an increasing giant component through time.

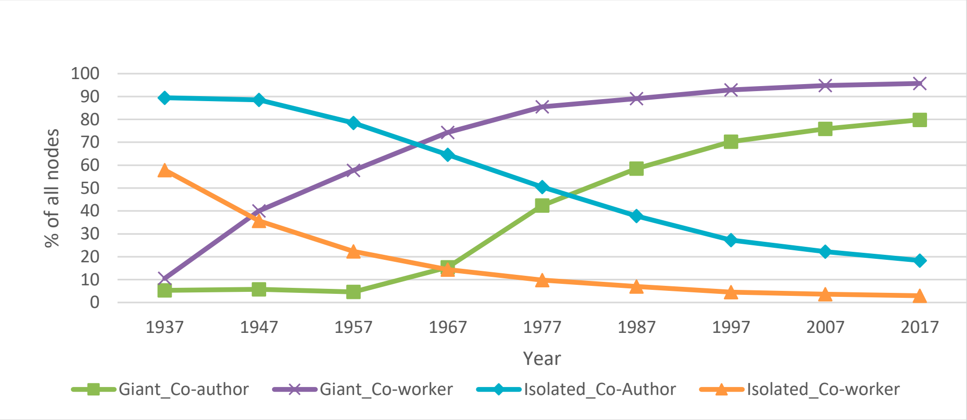


Figure 4. Size of Giant Component and Isolated Nodes

Conclusion

As the years pass, the Econometric Society is growing with newly elected Fellows. The Fellows establish more connections although the generational gap between Fellows widens. The Fellows publish more publications that strengthen their co-author links, and they continue to work in a few prestigious institutions that strengthen their co-worker links.

Acknowledgments

Part of this research project was conducted while I was visiting Queensland University of Technology. The visit is funded by the Scientific and Technological Research Council of Turkey (TÜBİTAK). I would like to thank Benno Torgler and Ho Fai Chan for their valuable comments.

References

Chan H.F & Torgler, B. (2012). Econometric Fellows and Nobel Laurates in Economics, Economics Bulletin, 32 (4), 3365-3377.

Goyal, S., van der Leij, M. & Moraga-González, J. L. (2006). Economics: An Emerging Small World, Journal of Political Economy, 114(2), 403-412.

Hamermesh, D.S & Schmidt, P. (2003). The Determinants of Econometric Society Fellows Elections, Econometrica, 71(1), 399-407.

Yuret, T. (2020). Co-worker networks: How closely are researchers who published in the top five economics journals related? Scientometrics, 124(3), 2301-2317

Topic Identification Index of Disruptive Technology Based on Patent Characteristics

Jiawei Zhang[1], Kaibiao Wu[2] and Yu Dong[3*]

[1] *zhangjiaw339@163.com* [2] *wukaibiao@mail.las.ac.cn* [3] *dongy@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing (China);
Department of Library, Information and Archives Management, School of Economics and Management,
University of Chinese Academy of Sciences, Beijing (China)

Abstract

Patent indicators are widely employed in the identification and prediction of disruptive technologies; however, extant research lacks analysis and evaluation of the applicability and effectiveness of these indicators. Therefore, it is necessary to investigate the patent indicators that best match disruptive technologies' salient characteristics. This study uses decision tree algorithms and correlation analysis to compare the characteristics presented by patent indicators found between disruptive and traditional technology fields to derive key indicators, and then verify the indicator research results to check the indicators' applicability and effectiveness. It shows that the changes in the number of patent applications can reveal the development and evolution of disruptive technologies to a certain extent; in terms of other indicators, number of patent family country, duration of patent examination, number of non-patents cited in patents, number of claims, number of patents cited by other patents, and the IPC (International Patent Classification) number are all key indicators. Besides, the index of patent text discrepancy can be used as an effective supplementary component to the above indicators. Finally, this study verifies the feasibility of the above indicators in identifying disruptive technologies in other patent collection fields, and these indicators can provide indicator references for subsequent related research.

Introduction

The rapid development of science and technology (S&T) is reshaping the world's competitive landscape. Among these elements, technological innovation has become the main focus. Bower and Christensen (1995) proposed the concept of disruptive technology for the first time by analyzing technological changes in small and medium-sized enterprises, thereby highlighting that disruptive technology can often replace existing mainstream technology in unexpected ways by not only opening up new markets but also forming new value systems. At present, many instances have revealed that in comparison to traditional or mainstream technologies, disruptive technologies have novel features, functions, development paths, etc., and can have a significant impact on technology paradigms, business models, and competitive situations. Therefore, disruptive technologies attract attention from the government and industry. For example, the Defense Advanced Research Projects Agency has developed a disruptive technological innovation management model to detect disruptive technologies (Cao et al. 2019). For companies, to maintain a competitive advantage in the market, they must rely on new technologies that have the potential to increase revenue (Taşkin, Adali, and Ersin 2004). In short, the identification and prediction of disruptive technologies have important societal, economic, and research significance.

Patent documents contain a wealth of technical information. The patent citation can reflect the ex post impact (Fleming 2001) or the market value (Gambardella, Harhoff, and Verspagen 2008) of the disruptive technology, and it can also be used to analyze the priori characteristics of the disruptive technology (Ahuja and Lampert 2001). Therefore, indicators designed for patent data are widely used to identify disruptive technologies or even predict their likely development (Momeni and Rost 2016). However, in the current practice of disruptive technology identification, the design of patent indicators is quite different. A unified standard for indicator evaluation has not yet been formed and the existing patent indicators mostly come from traditional qualitative technology identification such as foreseeable research. However, it does remain be further verified whether these indicators meet the essential characteristics of disruptive technologies. Besides, in current research, topic identification of disruptive technologies based on patent features is still relatively crude, and hence, it is vital to discuss how to adopt patent feature indicators that are more in line with disruptive technical features for identification. Given the above limitations, this study explores using patent data and machine learning methods to deeply excavate the the core characteristics and regular patterns of salient disruptive technologies from objective data and then build identification indicators of disruptive technology features that provide references for future follow-up work.

Related Work

Concept Development of Disruptive Technology

Initially, the concept of disruptive technology was aimed at managers who made investment or purchase decisions in the company and emphasized the impact of technology on developing existing products or markets (Christensen 1997). Then, the concept of "disruptive innovation" was proposed; it posits that the effects brought about by disruptive technologies are not limited to the technology itself but may also have a significant wider impact due to the introduction of new business models (Christensen and Raynor 2013). Although Christensen distinguished and explained the connotation of disruptive technology and innovation, these two concepts are often found to be used in combination without strict categorical boundaries in subsequent related studies (Dan and Chieh 2008). With disruptive technology's widespread in other fields, scholars have defined it from different perspectives, thereby expanding and enriching its significance (Table 1).

Table 1. Conceptual development of disruptive technology

Perspective	*Main points*	*Main viewpoints and representative scholars*
Market perspective	Emphasize the improvement and replacement of existing products	Disruptive technologies often provide a larger set of product performance combinations or differing performance implementation methods (Thomond and Lettice 2002; Huang and Sošić 2010).
	Pay attention to the relationship between disruptive technology and enterprise development	Defined from the perspective of latecomer companies occupying the market through disruptive technologies or in improving corporate competitiveness (Danneels 2004; Schmidt and Druehl 2008).
	Pay attention to customer needs, consumer behavior, etc.	Defined with consumer needs as the core and believes it can change market standards, consumer expectations, and destroy old markets or create new ones (Nagy, Schuessler, and Dubinsky 2016).
Technical perspective	Emphasize technological change and development	Disruptive technologies are combinations of existing technologies or brand-new technologies whose application will lead to a major paradigm shift (Kassicieh et al. 2002). Disruptive technology is a new dominant technology "derived" or "evolved" from established and technological systems, which can replace existing technologies and produce fundamental changes (Shaffer 2005).

A comprehensive analysis of the above concepts proposed by Christensen and subsequent scholars demonstrates that the definition of disruptive technology is mainly divided into two categories: one is based on Christensen's theory, focusing on the impact of technology on existing products or markets, and emphasizing the improvement or replacement of the existing market trends, and the second is to explain disruptive technologies from the perspective of the development of technology, focusing on the path of technological change and the related development, mode, and the significant impacts and changes thereby produced.

Table 2 Disruptive technology identification index based on patent characteristics

Technical features	*Identification index*	*Identification content*	*Identification features*
Technological breakthrough	Patent cited	The frequency of patent citations and changes	Reflects the influence of disruptive technology
	Patent technical topic	Topic similarity and difference of patent technology	Analyzes technical content such as patent abstracts and specifications, and evaluates technological breakthroughs
	Patent status	Including patent examination time, maintenance time, and litigation.	Reflects the cutting-edge breakthrough of technology
Technical Integration	IPC number	IPC quantity, IPC number combination, etc.	Cross-field or new field application of evaluation technology
	Number of patentees	Cross-field layout of related technologies owned by patentees	Assesses technology combination and integration capabilities
Technology development and diffusion	Number of applications	Number and changes of patent applications	Determines the period of technology
	Number of patents in the same family	Country count of patent family.	Reflects the market and technical value that can evaluate the feasibility of the technology
	Number of claims	Including the average number of claims, independent claims, and dependent claims	Reflects the feasibility of the technology entering the market
	Patent cited	The frequency of patent citations and changes	Reflects the influence, development, and diffusion of disruptive technologies
	Duration of patent examination	Time from patent application to authorization	Reflects the degree of recognition of patented technology

Literature Review of Disruptive Technology Identification Index Based on Patent Features

This study primarily focuses on the technological research perspective wherein research methods are mainly quantitative analysis, such as bibliometrics and patent analysis. Kostoff, Boylan, and Simons (2004) associate literature to identify disruptive technologies by integrating existing features and expert opinions for identification. After 2009, research on identification based on patent characteristics has gradually increased. Buchanan and Corken (2010) used patentee indicators to identify and evaluate whether a given technology can be said is disruptive. Dotsika and Watkins (2017) used keywords to construct a co-occurrence network to establish disruptive technology identification indicators, such as degree of centrality and intermediate centrality. Above all, as an essential source of information for technological innovation, patents are often used by scholars to explore, reveal, and assess the direction of or trend on technological development.

According to the definition of disruptive technology and relevant scholars' research, its main characteristics can be summarized into six aspects, namely, uncertainty, forward-looking, mutation, non-competitiveness, proliferation, and user value orientation (Brimley et al. 2013; Govindarajan and Kopalle 2006; Manyika et al. 2013). This study focuses on identifying disruptive technology topics from the perspective of patent characteristics and uses patent data to identify disruptive technologies' salient characteristics. Patent indicators, are generally divided into internal content feature and external feature indicators; the former includes the number of patent applications, patent citations, and patents of the same family, while the latter mainly involves the analysis of patent text content (Table 2).

Methods

This study's method is divided into four parts: a) the preliminary construction of the identification index system of disruptive technical characteristics, b) the identification of the critical indicators in the initially constructed index system, c) the construction of the evaluation index system, and d) the verification of the identification system (Figure 1).

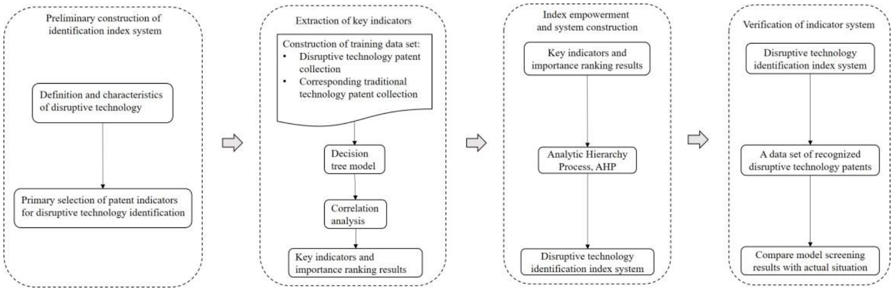

Preliminary construction of identification index system
Definition and characteristics of disruptive technology
Primary selection of patent indicators for disruptive technology identification
Extraction of key indicators
Construction of training data set:
• Disruptive technology patent collection
• Corresponding traditional technology patent collection
Decision tree model
Correlation analysis
Key indicators and importance ranking results
Index empowerment and system construction
Key indicators and importance ranking results
Analytic Hierarchy Process, AHP
Disruptive technology identification index system
Verification of indicator system
Disruptive technology identification index system
A data set of recognized disruptive technology patents
Compare model screening results with actual situation

Figure 1. Research framework

Preliminary Construction of Indicators

This study uses the decision tree algorithm to extract features that can distinguish between disruptive and traditional technology at the patent features level. The decision tree algorithm is a classic algorithm deployed in data mining and machine learning. It mainly addresses the classification problem through the layered reasoning of the tree structure. Its root node is the sample set, i.e., the patent data set in this study. Its internal nodes can be used to determine the attributes of the sample, i.e., the assessment of various indexes of patents. Its leaf node is the output of the decision result. The nodes in the decision tree are generated by the features found with the largest information gain. Therefore, the decision tree model can be used to obtain those indicators that have a greater influence on the identification of disruptive technology patents. Compared with other traditional classification algorithms, the results of the decision tree model are easy to interpret, and the classification rules can be directly extracted. Compared with deep learning models such as neural networks, decision tree models are more suitable for small and medium data sets and require fewer parameters, and the results are more in line with the human analysis of problems. Decision tree generation algorithms include ID3, C4.5, and CART (Du, Wang, and Gong 2011; Yi, Lu, and Liu 2011). In the process of decision tree model training, this study uses the Light Gradient Boosting Machine (LightGBM) model to perform a five-fold cross-validation on the training results to provide a foundation for selecting the best model results.

Extraction of Key Indicators

After a preliminary ranking of the indicators' importance, it is crucial to use correlation analysis to filter the indicators with a higher degree of correlation due to the correlation between these indicators. This study explicitly measures the Spearman correlation coefficient, the Pearson correlation coefficient, and the partial correlation coefficient of the above indicators. In the calculation, the subversion potential of related patents representing disruptive technologies is assigned a value of 1, while the subversion potential of related patents representing traditional technical fields is assigned a value of 0. Generally, the greater the value of the correlation coefficient, the stronger the correlation between the two given variables. As the above-mentioned patent indicators are derived from scientific metrological analysis, they will inevitably have the specific relevant characteristics. Therefore, merely measuring the relevance of each indicator from the correlation coefficient may cause individual deviations. Therefore, this study adopted the partial correlation analysis algorithm for index analysis and considered the potential subversion variable as the control variable to measure the actual degree of correlation between the indicators. In brief, this study used three indicators to measure the degree of correlation of the disruptive technology theme identification indicators and provide a screening criterion for the indicators obtained through the decision tree model.

Model Validation

After obtaining the corresponding indicators, the indicators' applicability needs to be verified. Specifically, after the decision tree model training input data, the best model is selected through a five-fold cross-validation of the model results. The model is then used to train the patent data set in a specific field (including the disruptive technology patent data set) to deduce whether it can discern it.

Experiments

Data

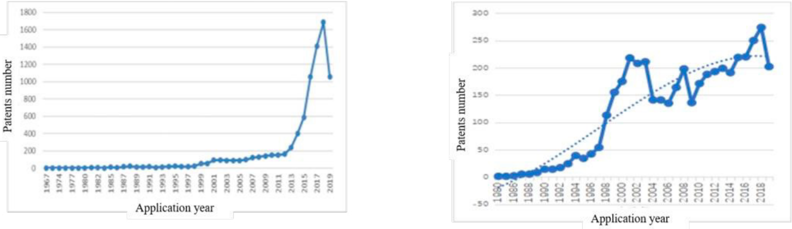
This study compared the disruptive technology's topic characteristics from the perspective of patent data by comparing the patent collections that have been identified as disruptive technologies and the parallel patent collections of traditional technologies. In terms of a specific selection of empirical fields, it selected biology as a valid disruptive technology paradigm and gene targeting technology as an exemplar of traditional technical field. The basis for choosing synthetic biology includes the fact that synthetic biology is a branch of biological sciences that only emerged in the 21st century. It originated from the reference to electronic circuits, which can imply that "humans can assemble organisms like assembled machines" becomes a reality (El Karoui, Hoyos-Flight, and Fletcher 2019). Synthetic biology is hailed as the third revolution in life sciences and has now been recognized as a disruptive technology. In recent years, synthetic biology has been widely used in many fields, including developing more effective vaccines and new drugs; improving drug production and development biological manufacturing, and biological management; and creating biosensors to detect toxic chemicals (Cameron, Bashor, and Collins 2014).

Synthetic biology is an interdisciplinary science, and its knowledge base and technical impetuses encompass the fields of chemistry, physics, information science, biological sciences, etc. It is not easy to correspond with the traditional technology of the development of synthetic biology. However, this study found that synthetic biology's hotspots are mainly gene editing to develop the programming of life systems, modular processing of metabolic pathways, and optimizing combinations between components, etc. Therefore, this study explored and selected the central corresponding traditional technology in this direction. Through literature research and expert consultation, it is found that gene targeting technology mainly uses screening markers and recombinases as the core to integrate external DNA into a specific position on the target cell's genome or site-specific mutations at predetermined target sites. It can be used to change a class of methods for cells' genetic characteristics (Capecchi 1989). Since its development, the new genome editing technology has the advantage of being a more efficient and straightforward technology, especially in the insertion and regulation of large-segment gene clusters. It has significantly improved the ability to change the cell genome accurately, which has led to the gradual development of traditional gene targeting technologies being supplemented (Meng and Ellis 2020). Based on this observation, this study aimed at researching synthetic biology-related technology patents and patents related to gene targeting technology and then construct a feature identification comparison to develop an indicator index.

After determining the empirical objects, Clarivate Analytics' Thomson Innovation (TI) database is used as a data source to procure data. This study focused on the themes of synthetic biology and gene targeting technology and used the database search function to search for relevant keywords to form the research data set. The time range of data collection is from the earliest application date retrieved in the database up to 2019. After obtaining the initial data, this study manually performed data deduplication, invalid data deletion, abnormal record screening, etc., to obtain the final patent data set, comprising 8195 patents related to synthetic biology from 1967 to 2019 and 4362 patents relate to gene targeting technology from 1980 to 2019.

Scientometric Analysis

The number of patent applications is often used to determine disruptive technologies' early neglected features. This study investigated the development trend of synthetic biology and gene targeting technology from the perspectives of changes in patent applications and changes in patent citations. By comparing these variables, it is noted that the development trend of the two is different, as shown in Figure 2.



Synthetic Biology

Gene targeting technology

Figure 2. Changes in the number of patents in synthetic biology and gene targeting technology

The development of synthetic biology techniques can be roughly divided into three stages based on the quantitative change trend:
Stage I: From 1967 to 2000, the number of synthetic biology-related patents was maintained at a low level. There are no more than 100 patent applications worldwide. It can be observed that its development has gone through a long process of incubation.
Stage II: From 2000 to 2007, the number of patent applications showed the first sudden change.
Stage III: Since 2008, the number of patents has increased exponentially. Moreover, it maintained a continuous upward trend and entered a rapid expansion stage, which saw a sudden marked upward trend in patent applications.

The changing trend of the number of patent applications for gene targeting technology can be fitted to the life cycle S curve: a) From 1980 to 1990, the number of patent applications increased gradually, which can be regarded as in the domain being in its infancy. This trend is different from the initial trend characteristics found in disruptive technologies. b) In 1994, gene targeting technology patents saw a significant increase for the first time and reached a peak in 2002, which was a period of rapid growth. c) From 2002 to 2007, the number of patent applications declined rapidly, reaching a trough in 2007; thereafter, the number of patent applications rose in fluctuations. Comprehensive qualitative research and expert consultation, in gene targeting technology, is currently in a mature development stage or perhaps even declining.

The changing trend of the patent applications number can reveal the development and evolution of disruptive technologies to a certain extent, including the initial long-term incubation process. When technological development shows a certain influence, the number of patents will undergo mutations. However, in the entire development process of disruptive technology, the number of patent applications has undergone more than one mutation. Therefore, only using the mutation characteristics of the number of patents to identify the embryonic period of disruptive technology will produce some errors, which need to be rectified through other analyses.

Index Construction for Disruptive Technology Identification

Through the previous literature research, 11 patent index features related to identifying disruptive technical topics are obtained. For the convenience of the following description, this study uses symbols to explain, the above indicators (Table 3).

Table 3 Symbol description of disruptive technology topic identification index

Symbol	*Indicators*	*Symbol*	*Indicators*
$F_{nCitPat}$	Number of cited patents	F_{nCla}	Number of claims
F_{nIPC}	IPC Number	$F_{nPatCit}$	Number of patents cited
$F_{nPatFam}$	Patent family number	$F_{nNonPatCit}$	Number of non-patents Cited in Patents
$F_{nPatFamCou}$	Number of patent family country	F_{PatSta}	Patent status (valid/invalid)
$F_{dPatExa}$	Duration of patent examination	F_{nPat}	Number of patentees
		$F_{PatTexTop}$	Patent textual topic

This study uses Python programming language to implement the decision tree classification model, takes the above-built data set and the patent indicators as the model's input items, and uses the technological subversion potential as the model's output items. During model training, the subversion potential of synthetic biology-related technology patents is assigned a value of 1, while the subversion potential of gene targeting technology patents is assigned a value of 0. In the decision tree model building process, first, we try to use only 10 external feature indicators. The accuracy of the cross-validation of training results is about 71% when using the Light GBM model. Then, after applying the TF-IDF (term frequency - inverse document frequency) algorithm and chi-square reduction to extract the patent title's content characteristics and considering text difference index into the model construction, we find that the accuracy of the model increases to 81%.

Using the decision tree model to analyze the results, the main indicators and their rankings that affect the score of technological disruption potential are obtained, which provides a basis for the subsequent construction of an indicator system for technology identification. As the decision tree model cannot distinguish between the importance of the patent's internal content feature and the external feature, only 10 external features' importance rankings are obtained (Table 4).

Table 4 Importance score of external features in disruptive technology topic identification

Indicator	$F_{nPatFamCou}$	$F_{dPatExa}$	$F_{nNonPatCit}$	F_{nCla}	$F_{nPatCit}$	F_{nIPC}	$F_{nCitPat}$	$F_{nPatFam}$	F_{PatSta}	F_{nPat}
Score	712	528	501	421	347	320	287	285	209	92

We used the correlation analysis algorithm in SPSS (Statistical Product and Service Solutions) v26.0 to analyze the index data of the above patent data set. Among them, the patent status index cannot be analyzed due to its unique data characteristics. The results reveal the following finding: a) Using Spearman correlation analysis can establish that the number of IPC, the number of citing patents, and the number of patents in the same family are more related to disruptive technologies' attributes and characteristics. b) Using Pearson correlation analysis can exhibit that the number of IPC numbers is strongly positively correlated with the number of patents in the same family. The number of cited patents is only slightly positively correlated with the number of patents in the same family. The number of citing patents is remotely positively correlated with the number of IPC numbers. The number of the patent family country is slightly positively correlated with the number of patent families. The correlation coefficients among other indicators are not high. c) When using partial correlation analysis to measure the correct degree of correlation between indicators, the results show a strong positive correlation between the number of patent family members and the number of IPCs. The number of cited patents is slightly positively correlated with the number of patent family members. The number of the patent family country is remotely positively correlated with the number of patent families (Table 5).

Combining the results of the correlation analysis of the above indicators, the number of IPCs has a strong correlation with the number of patents in the same family, and the number of cited patents has a marked impact influence on the number of patents in the same family and has a greater impact than the first six indicators. Therefore, this study selected the first six indicators as the key indicators of the identification index system: the number of patent family country, duration of patent examination, number of non-patents cited by patents, number of claims, number of patents cited by patents, and the IPC. After determining the criterion level and sub-criterion level of the index system and their indexes at all levels, the judgment matrix method is used to assign a relative weight to each index. Using the analytic hierarchy process to calculate the relative weight values between the above six external feature indicators, this study combines the important output results provided by the decision tree model and the above matrix scoring standards. After statistical collation, the following judgment matrix is obtained (Table 6).

After constructing the complete judgment matrix, the matrix operation is performed by the square root method to obtain the maximum eigenvalue of each indicator of the judgment matrix. The weight vectors of the matrix calculated by the relevant statistical algorithm are .46, .26, .14, .07, .04, .03 respectively, so the consistency test result is passed (CR = .067337161< .10). Therefore, it is judged that the index system constructed in this study is reasonable. The final technical feature identification system is thus obtained (Table 7).

Table 5 Correlation analysis results of disruptive technology identification indicators

	F_{nIPC}	F_{nCla}	F_{nPat}	$F_{nCitPat}$	$F_{nNonPatCit}$	$F_{nPatCit}$	$F_{nPatFam}$	$F_{dPatExa}$	$F_{nPatFamCou}$
Spearman correlation analysis									
Correlation coefficient	**-.112**	-.021	-0.19	.027	-0.87	**-1.35**	**-1.20**	**-.147**	**.174**
sig(bilateral)	.000	.018	.033	.003	.000	.000	.000	.000	.000
Pearson correlation analysis [The first line is Correlation coefficient, and the second line is sig(bilateral)]									
F_{nIPC}	1	-.064 .000	.209 .000	.113 .000	.095 .000	**.325** .000	**.700** .000	-0.28 0.12	.293 .000
F_{nCla}	-0.64 .000	1	-.125 .000	.034 .002	-.021 .060	-.002 .822	-0.96 .000	.053 .000	**-3.02** .000
F_{nPat}	.209 .000	-.125 .000	1	.109 .000	.128 .000	.220 .000	.254 .000	-0.26 .021	.215 .000
$F_{nCitPat}$	.113 .000	.034 .002	.109 .000	1	.131 .000	.065 .000	.095 .000	.031 .004	.134 .000
$F_{nNonPatCit}$	.095 .000	-.021 .060	.128 .000	.131 .000	1	.085 .000	.101 .000	.020 .067	.164 .000
$F_{nPatCit}$	.325 .000	-.002 .822	.220 .000	.065 .000	.085 .000	1	.500 .000	-.004 .723	.083 .000
$F_{nPatFam}$	.700 .000	-0.96 .000	-2.54 .000	.095 .000	-.101 .000	.500 .000	1	-0.41 .000	.337 .000
$F_{dPatExa}$	-.028 .012	.053 .000	-.026 .021	.031 .004	.020 .067	-.004 .723	-0.41 .000	1	-0.83 .000
$F_{nPatFamCou}$	.293 .000	-.302 .000	.215 .000	.134 .000	.164 .000	.083 .000	**.337** .000	-0.83 .000	1
Partial correlation analysis [The first line is Correlation coefficient, and the second line is sig(bilateral)]									
F_{nIPC}	1	-0.58 .000	-.207 .000	.097 .000	-0.78 .000	.271 .000	**.676** .000	-0.46 .000	**.310** .000
F_{nCla}	-.058 .000	1	-.155 .000	.032 .000	-.007 .427	.006 .480	-.100 .000	.098 .000	-.264 .000
F_{nPat}	.207 .000	-.155 .000	1	.094 .000	.118 .000	.130 .000	.262 .000	-.059 .000	.272 .000
$F_{nCitPat}$	.097 .000	.032 .000	.094 .000	1	.124 .000	.050 .000	.088 .000	.020 .026	.127 .000
$F_{nNonPatCit}$	.078 .000	-.007 .427	.118 .000	.124 .000	1	.070 .000	.096 .000	.035 .000	.161 .000
$F_{nPatCit}$	.271 .000	.006 .480	.130 .000	.050 .000	.070 .000	1	**.398** .000	.045 .000	.058 .000
$F_{nPatFam}$	.676 .000	-.100 .000	.262 .000	.088 .000	.096 .000	.398 .000	1	-0.74 .000	**.374** .000
$F_{dPatExa}$	-.046 .000	.098 .000	-0.59 .000	.020 .028	.035 .000	.045 .000	-.074 .000	1	-.118 .000
$F_{nPatFamCou}$	.310 .000	-.264 .000	.272 .000	.127 .000	.161 .000	.058 .000	.374 .000	-.118 .000	1

Table 6 Judgment matrix score

	$F_{nPatFamCou}$	$F_{dPatExa}$	$F_{nNonPatCit}$	F_{nCla}	$F_{nPatCit}$	F_{nIPC}
$F_{nPatFamCou}$	1	3	5	7	9	9
$F_{dPatExa}$	1/3	1	3	5	7	7
$F_{nNonPatCit}$	1/5	1/3	1	3	5	5
F_{nCla}	1/7	1/5	1/3	1	3	3
$F_{nPatCit}$	1/9	1/7	1/5	1/3	1	3
F_{nIPC}	1/9	1/7	1/5	1/3	1/3	1

Table 7 Technical feature identification index system

	Disruptive technological development						*Disruptive technology theme*
Indicator	$F_{nPatFamCou}$	$F_{dPatExa}$	$F_{nNonPatCit}$	F_{nCla}	$F_{nPatCit}$	F_{nIPC}	Patent text difference
index weight	*.46*	.26	.14	.07	.04	.03	-

Indicator Verification

This study uses the patent subset of the genome editing technology patent data set with the IPC number C07K001400 to verify the effectiveness of the above technical indicator system. Genome editing is a technology that uses a targeted modification of the genome to identify targets specifically for gene knockout and replacement. At present, three main genome editing technologies exit: artificial nuclease-mediated zinc finger nuclease technology, transcription activator-like effector nuclease technology, and RNA-mediated CRISPR-Cas9 nuclease technology. CRISPR-Cas9 technology is highly efficient and precise and has become a major milestone in developing genome-editing technology.

This study uses a subset of the IPC number C07K001400 in the genome editing technology patent. This patent subset is mainly related to the patented technology of peptides, which belongs to organic chemistry. A total of 288 related patents were retrieved in the TI database. According to the model's output, it is observed that 92.5% of the patents have attribute feature values higher than .5 points. The top five patents' IPC numbers are CN108610399A, KR2019113677A, CN108822217A, CN108822217A, WO2019185751A1 and CN107474129A. They are all related to the CRISPR-Cas9 system. It can be observed that the model can better identify disruptive technologies in patent concentrations.

Discussion and Conclusion

Following are the main study conclusions: The changing trend of patent applications' number can reveal the law of disruptive technologies' development and evolution to a certain extent; The effect of fusing the internal content features and external features of patent data is better than using only the external features to identify disruptive technologies; Six patent indicators are vital indicators according to decision tree model. These indicators are useful for subsequent use of patent data to identify subversion, while some of these indicators are somewhat different from traditional research. Then, this study further analyzes the significance of these indicators for the identification of disruptive technology:

a) Number of the patent family country: This indicator can reflect the patentee's recognition and confidence in the patent's development potential and market value. At present, in identification research, the number of patents in the same family is often used to measure this characteristic. However, this study revealed that the number of the patent family country may be more targeted.

b) Duration of patent examination: Different patents have different examination times due to their respective innovation development phases. Disruptive technology has the characteristics of being easily ignored or not being recognized in the early stage, which may affect the patent examiner's examination of the technology patent to a certain extent, thus affecting changes in the patent examination time. Currently, patent examination time has been used to research patent quality. This study highlighted that it can also be used in research related to disruptive technologies identification.

c) Number of non-patents cited in patents: The number of non-patents cited in patents reflects the relationship between technology and the S&T literature. When more non-patent literature is cited, it indicates that the technology patent has a sufficient scientific research foundation. This phenomenon is also in line with the long-term gestation and the phased outbursts of disruptive technologies.

d) Number of claims: The number of claims determines the scope of protection required by a patent, which is also a detailed description of the technical achievement characteristics. This indicator can reflect the characteristics of breakthroughs and the higher technological innovation of disruptive technologies.

e) Number of patents cited by patents: The quotations of patents to existing patents reflect the inheritance of technology or improve existing patents that show their innovation. In the overall index importance score, the back-cited features of patents are more important than the front-cited features, thereby illustrating the difference between disruptive technology and highly cited technology patents. Disruptive technologies may become highly cited patent technologies, but highly cited patents are not necessarily imbued with subversive potential. This aspect is also an identification standard that needs to be lucid and defined in identification research.

f) Number of the IPC: This aspect reflects the technical fields involved in the patent, such as reflecting the possibility of technological development or re-innovation of the technology in multiple technical fields and confirming the fact that disruptive technologies are new developments emerging from established and technical systems.

Research Limitations and Future Improvements

This study has the following limitations: (1) The construction of the index system in this article relies on the fields of synthetic biology and gene targeting technology, and the recognition effect in other fields remains to be verified. (2) In terms of extracting patent internal information, this study uses the TF-IDF algorithm, which ignores word position information, and thus, other models can be used to further extract patent text information.

As the development of disruptive technologies involves multiple aspects and types of data, this study will consider fusing patent data and market data to identify disruptive technologies in subsequent future research.

Acknowledgments

The first two authors contributed equally to this study. Yu DONG is corresponding author. This work is supported by Construction of Strategic Intelligence Research and Consulting System in the Field of S&T Policy (Special Program for the Construction of Literature Intelligence Capacity). The program number is E0290401.

References

Ahuja, Gautam, and Curba Morris Lampert. (2001). 'Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions', *Strategic management journal*, 22: 521-43.

Bower, Joseph L, and Clayton M Christensen. (1995). 'Disruptive technologies: catching the wave', *Harvard Business Review*,22: 43-53.

Brimley, Shawn, Ben FitzGerald, Kelley Sayler, and Peter Warren Singer. (2013). *Game changers: disruptive technology and US defense strategy*.

Buchanan, Ben, and Richard Corken. (2010). 'A toolkit for the systematic analysis of patent data to assess a potentially disruptive technology', *Intellectual Property Office United Kingdom, London*.

Cameron, D Ewen, Caleb J Bashor, and James J Collins. (2014). 'A brief history of synthetic biology', *Nature Reviews Microbiology*, 12: 381-90.

Cao, Xiaoyang, Yongjing Wei, Li Li, Ke Zhang, Hongbo Miao, Xiangchao An, and Anrong Liu. (2019). 'Enlightenment of disruptive technological innovation of DARPA', *Strategic Study of Chinese Academy of Engineering*, 20: 122-28.

Capecchi, Mario R. (1989). 'Altering the genome by homologous recombination', *Science*, 244: 1288-92.

Christensen, Clayton M. (1997). *The innovator's dilemma: when new technologies cause great firms to fail* (Harvard Business Review Press).

Christensen, Clayton, and Michael Raynor. (2013). *The innovator's solution: Creating and sustaining successful growth* (Harvard Business Review Press).

Dan, Yu, and Hang Chang Chieh. (2008). "A reflective review of disruptive innovation theory." *Portland International Conference on Management of Engineering & Technology*, 402-14. IEEE.

Danneels, Erwin. (2004). 'Disruptive technology reconsidered: A critique and research agenda', *Journal of product innovation management*, 21: 246-58.

Dotsika, Fefie, and Andrew Watkins. (2017). 'Identifying potentially disruptive trends by means of keyword network analysis', *Technological Forecasting and Social Change*, 119: 114-27.

Du, Ming, Shu Mei Wang, and Gu Gong. (2011). "Research on decision tree algorithm based on information entropy." In *Advanced Materials Research*, 732-37. Trans Tech Publ.

El Karoui, Meriem, Monica Hoyos-Flight, and Liz Fletcher. (2019). 'Future Trends in Synthetic Biology–A report', *Frontiers in bioengineering and biotechnology*, 7: 175.

Fleming, Lee. (2001). 'Recombinant uncertainty in technological search', *Management science*, 47: 117-32.

Gambardella, Alfonso, Dietmar Harhoff, and Bart Verspagen. (2008). 'The value of European patents', *European Management Review*, 5: 69-84.

Govindarajan, Vijay, and Praveen K Kopalle. (2006). 'The usefulness of measuring disruptiveness of innovations ex post in making ex ante predictions', *Journal of product innovation management*, 23: 12-18.

Huang, Xiao, and Greys Sošić. (2010). 'Analysis of industry equilibria in models with sustaining and disruptive technology', *European Journal of Operational Research*, 207: 238-48.

Kassicieh, Suleiman K, Steven T Walsh, John C Cummings, Paul J McWhorter, Alton D Romig, and W David Williams. (2002). 'Factors differentiating the commercialization of disruptive and sustaining technologies', *IEEE transactions on engineering management*, 49: 375-87.

Kostoff, Ronald N, Robert Boylan, and Gene R Simons. (2004). 'Science and Technology Test Mining: Disruptive Technology Roadmaps', *Technological Forecasting and Social Change*, 71: 141-59.

Manyika, James, Michael Chui, Jacques Bughin, Richard Dobbs, Peter Bisson, and Alex Marrs. (2013). "Disruptive technologies: Advances that will transform life, business, and the global economy." In.: McKinsey Global Institute San Francisco, CA.

Meng, Fankang, and Tom Ellis. (2020). 'The second decade of synthetic biology: 2010–2020', *Nature Communications*, 11: 1-4.

Momeni, Abdolreza, and Katja Rost. (2016). 'Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling', *Technological Forecasting and Social Change*, 104: 16-29.

Nagy, Delmer, Joseph Schuessler, and Alan Dubinsky. (2016). 'Defining and identifying disruptive innovations', *Industrial Marketing Management*, 57: 119-26.

Schmidt, Glen M, and Cheryl T Druehl. (2008). 'When is a disruptive innovation disruptive?', *Journal of product innovation management*, 25: 347-69.

Shaffer, Alan R. (2005). "Disruptive Technology: An Uncertain Future." *DIRECTOR DEFENSE RESEARCH AND ENGINEERING WASHINGTON DC PLANS AND PROGRAMS*.

Taşkin, Harun, Mehmet Riza Adali, and Eren Ersin. (2004). 'Technological intelligence and competitive strategies: an application study with fuzzy logic', *Journal of Intelligent Manufacturing*, 15: 417-29.

Thomond, Peter, and Fiona Lettice. (2002). "Disruptive innovation explored." *Cranfield University, Cranfield, England. Presented at: 9th IPSE International Conference on Concurrent Engineering: Research and Applications* .

Yi, Weiguo, Mingyu Lu, and Zhi Liu. (2011). 'Multi-valued attribute and multi-labeled data decision tree algorithm', *International Journal of Machine Learning and Cybernetics*, 2: 67-74.

Two New Field Normalization Indicators Considering the Reliability of Citation Time Window: Some Theoretical Considerations

Lihua Zhang[1] and Xing Wang[2]*

[1] *happy2004zlh@163.com*
Shanxi University of Finance & Economics, School of Information, 696 Wucheng Road, 030006 Taiyuan (China)

[2] *wangxing@sjtu.edu.cn* (* **Corresponding Author**)
Shanxi University of Finance & Economics, School of Information, 696 Wucheng Road, 030006 Taiyuan (China)

Abstract

Field normalization indicators are not reliable when a short citation time window is used. To solve this problem, two new field normalization indicators called the Weighted Category Normalized Citation Impact (WCNCI) and Total WCNCI (TWCNCI) were proposed by Wang and Zhang (Wang & Zhang, 2020; Wang & Zhang, 2021). The main idea of the two new normalization indicators is that a weighting factor, which is calculated as the correlation coefficient between citation counts of papers in the given short citation time window and in the fixed-length long window, is introduced to represent the degree of reliability of normalized citations. In this study, we conducted a theoretical analysis of the two new indicators. We first compared WCNCI with the crown indicator and new crown indicator in terms of how to handle recent publications. We found that WCNCI has the advantages of both the crown and new crown indicators. Furthermore, we also proved that WCNCI and TWCNCI have the consistency property, and that WCNCI simultaneously has the homogeneous normalization property. Additionally, we discussed how to handle overlapping fields when calculating WCNCI and TWCNCI.

Introduction

The citation count is a significantly important indicator used to measure the impact of publications in research evaluation. In some fields, the average number of citations per publication is much higher than those in other fields because of the inherent differences in citation practice among fields (Waltman et al., 2011b). For example, the citation counts of papers in Molecular Biology & Genetics are generally higher than those in Mathematics. The Essential Science Indicator (ESI) updated in January 2021 reported that the average number of citations per paper in Molecular Biology & Genetics was 24.58, whereas it was only 4.89 in Mathematics (Clarivate Analytics, 2021). Therefore, the citation counts of papers from different fields should not be compared directly. Field normalization methods are recommended in the guiding principles for research evaluation in the Leiden Manifesto (Hicks et al., 2015).

The literature on field normalization is extensive. The main methods include the mean-based method (Lundberg, 2007; Waltman et al., 2011b), *z*-score method (Vaccario et al., 2017), percentile rank method (Bornmann, Leydesdorff, & Wang, 2013; Bornmann & Williams, 2020; Leydesdorff & Bornmann, 2011), reverse engineering method (Radicchi & Castellano, 2012), citing-side method (Leydesdorff & Opthof, 2010; Waltman & van Eck, 2013; Zitt & Small, 2008), and method that combines the citing-side and percentile methods (Bornmann, 2020).

An important problem to solve is that field normalization indicators are not reliable when a short citation time window is used. Recent papers do not have sufficient time to be cited, and their citation counts are not as reliable as those of papers published many years ago. Based on an investigation of Web of Science journal papers published in 1980, Wang (Wang, 2013) found that the correlation coefficient between the citation counts in a 2-year citation time window and those in a more reliable citation window of 31 years was only 0.592 for all fields, and the correlation coefficients were even lower in the fields of Engineering Technology and Mathematics, dropping to 0.466 and 0.386, respectively. According to Nederhof et al. (Nederhof, Van Leeuwen, & Clancy, 2012), longer citation time windows yielded more favorable results than shorter windows, that is, the longer the citation time window, the more reliable the field normalization indicators. However, research evaluation practice is usually conducted in a short time period and cannot wait for decades. To address this problem, Wang and Zhang proposed the following solution (Wang & Zhang, 2020).

The normalization citation count of each paper is assigned a weighting factor to represent its degree of reliability. The weighting factor is calculated as the correlation coefficient between the citation counts in the short citation time window and those in a fixed reliable long citation time window (e.g., 31 years). The new weighted indicator at the aggregation levels (e.g., at the country level, institution level, or research group level) is called Weighted Category Normalized Citation Impact (WCNCI) (Wang & Zhang, 2020) and is an average performance indicator. Average performance indicators are scale free and cannot measure the total performance of a set of publications. Because of this, Wang and Zhang proposed a total performance indicator called Total WCNCI (TWCNCI), which equals the number of publications multiplied by the WCNCI of these publications (Wang & Zhang, 2021).

In this paper, we present a theoretical analysis of WCNCI and TWCNCI. We first compare WCNCI with CPP/FCSm (De Bruin, Kint, & Moed, 1993; Moed, Debruin, & Vanleeuwen, 1995) and Mean Normalized Citation Score (MNCS) (Waltman et al., 2011b) in terms of how to handle recent publications. Furthermore, we particularly pay attention to the consistency property of WCNCI and TWCNCI. Finally, we discuss how to handle overlapping fields when calculating WCNCI and TWCNCI.

Definition of indicators

In this section, we provide formal definitions of the six indicators: (1) average performance indicators: CPP/FCSm, MNCS and WCNCI; and (2) total performance indicators: CPP/FCSm×n (Waltman et al., 2011b), Total Normalized Citation Score (TNCS) (Waltman et al., 2011b) and TWCNCI.

CPP/FCSm is called the “crown indicator” and has been used for more than 20 years, whereas MNCS is known as the new crown indicator and adopts a different normalization mechanism.

CPP/FCSm is defined as

$$CPP\,/\,FCSm = \frac{\sum_{i=1}^{n} c_i / n}{\sum_{i=1}^{n} e_i / n} = \frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} e_i} \qquad (1)$$

and MNCS is defined as

$$MNCS = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{e_i}, \tag{2}$$

where c_i is the raw citation count of publication i. e_i is the expected citation count of publication i, which equals the average citation counts of all publications published in the same field, same year and with the same document type as publication i. n is the total number of publications in a publication set.

The CPP/FCSm indicator is the average performance ratio, whereas MNCS is the average of the individual performance ratios. CPP/FCSm and MNCS are both average indicators that measure the average performance of a set of publications. They cannot measure the total performance of a set of publications. Thus, we use two total performance indicators: CPP/FCSm × n and TNCS.

CPP/FCSm × n is defined as

$$CPP\,/\,FCSm\times n = \frac{\sum_{i=1}^{n}c_i}{\sum_{i=1}^{n}e_i/n} = \frac{n\sum_{i=1}^{n}c_i}{\sum_{i=1}^{n}e_i} \tag{3}$$

and TNCS is defined as

$$TNCS = MNCS\times n = \sum_{i=1}^{n}\frac{c_i}{e_i}. \tag{4}$$

The normalized citation indicator may not be sufficiently reliable when a short citation time window is used because the citation counts of recent papers are not as reliable as those of older papers. To solve this problem, Wang and Zhang introduced a new indicator called WCNCI (Wang & Zhang, 2020), which is defined as

$$WCNCI = \frac{1}{n}\sum_{i=1}^{n}w_i\frac{c_i}{e_i}, \tag{5}$$

where w_i is the weight of publication i, which represents the degree of reliability of the normalized citation count of publication i. w_i is equal to the correlation coefficient between citation counts of publications in the given short citation time window and those in a fixed long citation time window. For instance, if the correlation coefficient between the citation counts of papers published 2 years ago and 31 years ago in chemistry is 0.55, the normalized citation count of a chemistry paper with a citation window of 2 years should be multiplied by 0.55 to obtain its reliable scientific impact. The shorter (longer) the time window after publication, the lower (higher) the correlation coefficient and degree of reliability.

Similar to CPP/FCSm and MNCS, WCNCI also has a corresponding total performance indicator TWCNCI as follows:

$$TWCNCI = WCNCI\times n = \sum_{i=1}^{n}w_i\frac{c_i}{e_i}. \tag{6}$$

How to handle recent publications

How to handle recent publications is an important issue in the process of field normalization. It is not sufficiently reliable to calculate the field normalization indicator with a short citation time window, such as 2 years. Recent papers have not had sufficient time to accumulate citations after publication. However, using a short citation time window to evaluate the recent research performance of journals, researchers, institutions and countries is generally inevitable in research evaluation practice. For example, university rankings are usually published once a year and they generally use recent publications for the paper indicators. Scientific research administrators and policymakers also usually cannot wait for decades to evaluate the research impact of publications in actual research evaluation practice.

CPP/FCSm can be regarded as a type of weighted version:

$$CPP/FCSm = \frac{1}{n}\sum_{i=1}^{n} w_i \frac{c_i}{e_i}, \tag{7}$$

where w_i is given by

$$w_i = \frac{e_i}{\sum_{j=1}^{n} e_j / n}. \tag{8}$$

The weight w_i in CPP/FCSm implicitly allocates more (less) weight to older (more recent) papers because older (more recent) papers naturally have a relatively higher (lower) *e* value. However, w_i in CPP/FCSm simultaneously allocates more (less) weight to papers from fields with a higher (lower) expected number of citations, which is unreasonable and contrary to the original intention of field normalization, which is to eliminate field differences in citation counts and ensure that normalized citations are comparable across fields (Wang & Zhang, 2020). Unlike CPP/FCSm, MNCS gives the same weight to all publications; that is, recent publications play the same role as older publications. According to the opinion of Waltman et al. (Waltman et al., 2011a), the calculation of the MNCS indicator could leave out recent publications published less than 1 year to overcome quite a significant amount of noise in the indicator caused by recent publications.

However, leaving out recent publications may result in the loss of some useful information. Furthermore, leaving out recent publications cannot essentially solve the problem that the field normalization indicator is not reliable when a short citation time window is used because the degree of reliability of citation counts from different citation time windows is different; even the citation count of a paper published 10 years is not reliable compared with a paper published 30 years. Wang and Zhang proposed a new solution to this problem (Wang & Zhang, 2020). The normalization citation count of each paper is assigned a weighting factor to represent its degree of reliability. The weighting factor is calculated as the correlation coefficient between the citation counts in the short citation time window and those in a fixed reliable long citation time window. The shorter (longer) the citation time window, the lower (higher) the correlation coefficient and degree of reliability.

To compare the difference in handling recent publications between CPP/FCSm, MNCS and WCNCI, we consider the following fictitious example. Suppose there are two research units: research unit A and research unit B. Both research units are active in field X and field Y.

Research unit A publishes 10 publications during 10 years in field X and field Y, and so does research unit B. The characteristics of the datasets used to compare the three indicators are reported in Table 1. The numbers in parentheses represent the citation count and expected citation count of a publication. For example, (1, 0.1) means that a publication has been cited once and the average citation counts of all peer publications is 0.1. Peer publications refer to publications published in the same year and field as the publication investigated.

Table 1. Characteristics of the datasets used to compare CPP/FCSm, MNCS and WCNCI.

		Publishing age									
		1	2	3	4	5	6	7	8	9	10
Research Unit A	Field X	(1, 0.1)	—	(2, 3)	—	(5, 4)	—	(7, 6)	—	(8, 7)	—
	Field Y	(2, 0.15)	—	(5, 4)	—	(8, 7)	—	(12, 10)	—	(15, 12)	—
Research Unit B	Field X	(2, 0.1)	—	(10, 3)	—	(15, 4)	—	(20, 6)	—	(25, 7)	—
	Field Y	(3, 0.15)	—	(15, 4)	—	(24, 7)	—	(30, 10)	—	(35, 12)	—

Note: Research units A and B do not publish any papers in some years. This case is indicated by "—."

Weighting factors are necessary when WCNCI is calculated, and they are listed in Table 2. Five citation time windows are used: 1 year, 3 years, 5 years, 7 years and 9 years.

Table 2. Weighting factors of field X and field Y in different citation time windows.

	Citation time window				
	1 year	3 years	5 years	7 years	9 years
Field X	0.10	0.30	0.50	0.70	0.80
Field Y	0.15	0.40	0.65	0.80	0.95

Table 3 presents the values of the three indicators WCNCI, CPP/FCSm and MNCS calculated for research units A and B using two different approaches: including all publications and leaving out publications published no more than 1 year. The fourth column and sixth column show the differences between the two different approaches for the same indicator and research unit.

Table 3. Changes after leaving out recent publications for the three indicators.

Indicator	*Methods*	*Research unit A*	*Difference A*	*Research unit B*	*Difference B*
CPP/FCSm	All papers	1.22	0.05	3.36	0.08
	1-year papers excluded	1.17		3.28	
MNCS	All papers	3.24	2.11	6.71	3.32
	1-year papers excluded	1.13		3.39	
WCNCI	All papers	0.89	0.15	2.20	0.08
	1-year papers excluded	0.74		2.12	

As seen in Table 3, the values of the MNCS indicator change more significantly than those of CPP/FCSm and WCNCI after recent publications are left out. This demonstrates that MNCS is more easily influenced by recent publications than the other two indicators. Clearly, both CPP/FCSm and WCNCI have advantages over MNCS in terms of handling recent publications. Despite this, CPP/FCSm is not recommended because it allocates more (less) weight to papers from fields with a higher (lower) expected number of citations, which is unfair because different fields should be treated equally.

Consistency of indicators

In this section, we prove the consistency of WCNCI and TWCNCI. Consistency is a mathematical property that some indicators have and others do not.

We define some mathematical notation first. A publication can be expressed as an ordered triple (c, e, w), where c and e denote the actual citation count and expected citation count, respectively. The expected citation count of a publication equals the average citation counts of all publications published in the same field, in the same year and with the same publication type. w is the weight of a publication that represents the degree of reliability of the normalization citation count. c is a positive integer, whereas e and w are both positive real numbers. P is defined as the set of all ordered triples (c, e, w). The set including n publications can be represented by a multiset S = { (c_1, e_1, w_1), (c_2, e_2, w_2), … (c_n, e_n, w_n) }, where (c_1, e_1, w_1), (c_2, e_2, w_2), … (c_n, e_n, w_n)$\in P$. $\sum$ is the set of all non-empty multisets S.

WCNCI is an average performance indicator and TWCNCI is a total performance indicator. First, we prove the consistency of TWCNCI.

Definition 1. Let f_T denote a bibliometric indicator of the total performance of a set of publications. f_T is said to have the consistency property if

$$f_T(S_1) \geq f_T(S_2) \Leftrightarrow f_T(S_1 \cup \{(c,e,w)\}) \geq f_T(S_2 \cup \{(c,e,w)\}) \tag{9}$$

for all S_1, $S_2 \in \sum$ and all $(c, e, w) \in P$.

Definition 1 means that an indicator of total performance is consistent if adding the same publication to two different sets of publications never changes the way in which the indicator ranks the sets of publications relative to each other.

We assume that there are n_1 publications in S_1 and n_2 publications in S_2. According to the definition of TWCNCI,

$$\begin{aligned} f_T(S_1) \geq f_T(S_2) &\Leftrightarrow \sum_{i=1}^{n_1} w_i \frac{c_i}{e_i} \geq \sum_{j=1}^{n_2} w_j \frac{c_j}{e_j} \Leftrightarrow \left(w\frac{c}{e} + \sum_{i=1}^{n_1} w_i \frac{c_i}{e_i} \right) \geq \left(w\frac{c}{e} + \sum_{j=1}^{n_2} w_j \frac{c_j}{e_j} \right) \\ &\Leftrightarrow f_T(S_1 \cup \{(c,e,w)\}) \geq f_T(S_2 \cup \{(c,e,w)\}) \end{aligned} \tag{10}$$

TWCNCI has the consistency property.

We now consider the indicator WCNCI, which measures the average performance of a set of publications. For an indicator of average performance, we use a slightly different definition of consistency.

Definition 2. Let f_A denote a bibliometric indicator of the average performance of a set of publications. f_A is said to have the consistency property if

$$f_A(S_1) \geq f_A(S_2) \Leftrightarrow f_A(S_1 \cup \{(c,e,w)\}) \geq f_A(S_2 \cup \{(c,e,w)\}) \tag{11}$$

for all S_1, $S_2 \in \sum$ such that $|S_1| = |S_2|$, and all $(c, e, w) \in P$.

According to Definition 2, an average performance indicator is consistent if adding the same publication to two different but equally large sets of publications never changes the way in which the indicator ranks the sets of publications relative to each other.

We assume that there are *n* publications in both S_1 and S_2. According to the definition of WCNCI,

$$\begin{aligned} & f_A(S_1) \geq f_A(S_2) \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} w_i \frac{c_i}{e_i} \geq \frac{1}{n}\sum_{j=1}^{n} w_j \frac{c_j}{e_j} \Leftrightarrow \sum_{i=1}^{n} w_i \frac{c_i}{e_i} \geq \sum_{j=1}^{n} w_j \frac{c_j}{e_j} \\ & \Leftrightarrow \frac{1}{n+1}\left(w\frac{c}{e} + \sum_{i=1}^{n} w_i \frac{c_i}{e_i} \right) \geq \frac{1}{n+1}\left(w\frac{c}{e} + \sum_{j=1}^{n} w_j \frac{c_j}{e_j} \right) \\ & \Leftrightarrow f_A(S_1 \cup \{(c,e,w)\}) \geq f_A(S_2 \cup \{(c,e,w)\}) \end{aligned} \tag{12}$$

WCNCI has the consistency property.

The average performance indicator has another property called homogeneous normalization, which was discussed by Waltman (Waltman et al., 2011b).

Definition 3. Let f_A denote a bibliometric indicator of the average performance of a set of publications. f_A is said to have the homogeneous normalization property if

$$f_A(S) = \frac{\sum_{i=1}^{n} w_i c_i}{ne} \tag{13}$$

for all S = { (c_1, e_1, w_1), (c_2, e_2, w_2), … (c_n, e_n, w_n) } $\in \sum$ and $e_1 = e_2 = ... = e_n = e$.

The homogeneous normalization property aims to describe homogeneous sets of publications. In this paper, a homogeneous set of publications means that all publications belong to the same field and were published in the same year. It is not difficult to confirm that WCNCI has the homogeneous normalization property.

To summarize, WCNCI and TWCNCI both have the consistency property, whereas WCNCI simultaneously has the homogeneous normalization property.

How to handle overlapping fields

In this section, we focus on how to calculate the values of WCNCI and TWCNCI in the case of overlapping fields.

Because some journals in the Web of Science database may belong to more than one subject category, if we calculate WCNCI and TWCNCI based on the subject categories in the Web of Science database, we will encounter the problem of handling overlapping fields, that is, when a publication belongs to more than one subject category. Therefore, we need to consider how WCNCI and TWCNCI should be calculated.

We use a fictitious example as an illustration. Suppose there are only three fields: field X, field Y and field Z. Five publications are distributed across the three fields. For each publication, the field it belongs to, number of citations, years since publication and weight representing the degree of reliability of the normalization citation count are listed in Table 4. Notice that publication 5 simultaneously belongs to field X and field Y.

Table 4. Five publications and the fields to which they belong.

	Field	*Citations*	*Publishing age*	*Weight*
Publication 1	X	2	3	0.15
Publication 2	X	3	3	0.15
Publication 3	Y	6	1	0.10
Publication 4	Z	10	5	0.50
Publication 5	X and Y	4	4	0.25

The weighting factors of the three fields are listed in Table 5. Publication 1 belongs to field X, and its weight is 0.15 because the correlation coefficient between the citation counts in the 3-year citation time window and those in the fixed long citation window (31 years) in field X equals 0.15. The weights of publications 2, 3 and 4 equal 0.15, 0.10 and 0.50, respectively. Half of publication 5 belongs to field X and half to field Y. Its weight equals the average weight of field X and field Y, that is, 0.25.

Table 5. Weighting factors of fields X, Y and Z.

	Citation time window				
	1 year	2 years	3 years	4 years	5 years
Field X	0.05	0.10	0.15	0.20	0.30
Field Y	0.10	0.20	0.25	0.30	0.40
Field Z	0.20	0.24	0.32	0.40	0.50

Now we consider calculating the expected citation counts of the five publications. Publications 1, 2, 3 and 4 belong to one field, and it is relatively easy to calculate the expected citation counts. Publication 5 belongs to fields X and Y, and its citations are divided equally between fields X and Y. Both publications 1 and 2 belong to field X, and therefore their expected citation counts equal the average citation counts of all publications published in field X. The expected citation counts of publications 1 and 2 are given by

$$e_1 = e_2 = \frac{2+3+4/2}{1+1+1/2} = \frac{14}{5}. \tag{14}$$

As can be seen, half of the citation count of publication 5 is used in Equation (14) because publication 5 is divided equally between fields X and Y. The expected citation count of publication 3 equals

$$e_3 = \frac{6+4/2}{1+1/2} = \frac{16}{3}, \tag{15}$$

where publication 5 is also divided equally between fields X and Y.

The expected citation count of publication 4 equals 10 because field Z consists of only one publication that is cited 10 times, and e_4 =10.

It is slightly complicated to calculate the expected citation count of publication 5. Two methods exist, which are the arithmetic average and harmonic average of e_1 and e_3 when publication 5 simultaneously belongs to field X and field Y. When the arithmetic average is adopted,

$$e_5 = \frac{14/5 + 16/3}{2} = \frac{244}{15}. \tag{16}$$

Calculating the values of WCNCI and TWCNCI for the set of five publications, we obtain

$$WCNCI = \frac{1}{5}\left(0.15 \times \frac{2}{14/5} + 0.15 \times \frac{3}{14/5} + 0.1 \times \frac{6}{16/3} + 0.5 \times \frac{10}{10} + 0.25 \times \frac{4}{244/15}\right) \approx 0.19 \tag{17}$$

$$TWCNCI = 5 \times 0.19 = 0.95. \tag{18}$$

The other method used to calculate the expected citation count is the harmonic average. Then we have

$$e_5 = \frac{2}{\frac{1}{14/5} + \frac{1}{16/3}} = \frac{224}{61}, \tag{19}$$

which yields

$$WCNCI = \frac{1}{5}\left(0.15 \times \frac{2}{14/5} + 0.15 \times \frac{3}{14/5} + 0.1 \times \frac{6}{16/3} + 0.5 \times \frac{10}{10} + 0.25 \times \frac{4}{224/61}\right) \approx 0.23 \tag{20}$$

$$TWCNCI = 5 \times 0.23 = 1.15. \tag{21}$$

According to Waltman et al. (Waltman et al., 2011b), MNCS has the nature of a baseline. If the value of the MNCS indicator of a research unit equals one, this means that the citation impact of the research unit equals the world average. If the value of the MNCS indicator is greater (less) than one, this means that the citation impact of the research unit is above (below) the world average. Waltman et al. (Waltman et al., 2011b) used the arithmetic average and harmonic average approach respectively to calculate the expected citation counts of publications published in more than one field. They found that the harmonic average approach could ensure that the baseline of MNCS always equals one. They therefore argued that the most appropriate approach to deal with overlapping fields is the harmonic average. The WCNCI indicator also has the nature of a baseline. However, the baseline of WCNCI is not equal to one, as is MNCS, but equals the average of all publications' weights. In our example above, the baseline of WCNCI could be calculated as

$$WCNCI_{baseline} = \frac{1}{5}(0.15 + 0.15 + 0.1 + 0.5 + 0.25) = 0.23. \tag{22}$$

We adopt both the arithmetic average and harmonic average to calculate the average citation impact of different fields in Equations (17) and (20). From the perspective of the baseline nature of the WCNCI indicator, it is obvious that the harmonic average approach is more appropriate for dealing with the problem of overlapping fields. This finding coincides with that of Waltman et al. (Waltman et al., 2011b). However, the arithmetic average approach provides a clear physical meaning and intuitive interpretation of the citation impact of a research unit. The two average approaches above provide opposite choices. In the future, more studies need to be conducted on how to select an appropriate average approach between the arithmetic average and harmonic average in the case of handling overlapping fields.

Discussion and conclusion

There have been several arguments about whether the CPP/FCSm indicator or MNCS indicator is more appropriate. Generally, academics prefer MNCS because it has a clear physical meaning. However, MNCS is not perfect in dealing with publications with different citation time windows. The citation impact of a recent publication is unreliable because recent publications have not had sufficient time to be cited; that is, recent publications and publications published many years ago should be treated in a different way when calculating the normalized citation impact. This is the limitation of the new crown indicator MNCS.

CPP/FCSm can be regarded as a type of weighted version of the MNCS indicator. It implicitly gives more (less) weight to older (more recent) publications. However, CPP/FCSm simultaneously gives more weight to publications from fields that have a higher expected number of citations, which is unfair because all fields should be treated equally.

WCNCI and TWCNCI retain the advantages and avoid the disadvantages of CPP/FCSm and MNCS. The two new field normalization indicators directly give more (less) weight, which is calculated as the correlation coefficient between citation counts of publications in the given short citation time window and those in a fixed long citation time window, to older (more recent) publications.

In this paper, we proved that both WCNCI and TWCNCI have the consistency property and that WCNCI simultaneously has the homogeneous normalization property. We also discussed the problem of handling overlapping fields when calculating WCNCI and TWCNCI. The arithmetic average approach and harmonic average approach are two commonly used methods. Each has pros and cons. There is no consensus on which one is better, so more studies need to be conducted in the future.

Acknowledgments

We thank Zhihui Zhang, PhD, from CNKI for his valuable suggestions on earlier draft of this paper.

References

Bornmann, L. (2020). How can citation impact in bibliometrics be normalized? A new approach combining citing-side normalization and citation percentiles. *Quantitative Science Studies, 1*(4), 1553-1569.

Bornmann, L., Leydesdorff, L., & Wang, J. (2013). Which percentile-based approach should be preferred for calculating normalized citation impact values? An empirical comparison of five approaches including a newly developed citation-rank approach (P100). *Journal of Informetrics, 7*(4), 933-944.

Bornmann, L., & Williams, R. (2020). An evaluation of percentile measures of citation impact, and a proposal for making them better. *Scientometrics, 124*(2), 1457-1478.

Clarivate Analytics. (2021). Essential Science Indicators.

De Bruin, R. E., Kint, A., Luwel, M., & Moed, H. F. (1993). A study of research evaluation and planning: The University of Ghent. *Research Evaluation, 3*(1), 25-41.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. [Note]. *Nature, 520*(7548), 429-431.

Leydesdorff, L., & Bornmann, L. (2011). Integrated impact indicators compared with impact factors: an alternative research design with policy implications. *Journal of the American Society for Information Science and Technology, 62*(11), 2133-2146.

Leydesdorff, L., & Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology, 61*(11), 2365-2369.

Lundberg, J. (2007). Lifting the crown-citation z-score. *Journal of Informetrics, 1*(2), 145-154.

Moed, H. F., Debruin, R. E., & Vanleeuwen, T. N. (1995). New bibliometric tools for the assessment of national research performance - database description, overview of indicators and first applications. *Scientometrics, 33*(3), 381-422.

Nederhof, A. J., Van Leeuwen, T. N., & Clancy, P. (2012). Calibration of bibliometric indicators in space exploration research: a comparison of citation impact measurement of the space and ground-based life and physical sciences. *Research Evaluation, 21*(1), 79-85.

Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *Plos One, 7*(3), e33833.

Vaccario, G., Medo, M., Wider, N., & Mariani, M. S. (2017). Quantifying and suppressing ranking bias in a large citation network. *Journal of Informetrics, 11*(3), 766-782.

Waltman, L., & van Eck, N. J. (2013). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics, 7*(4), 833-849.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011a). Towards a new crown indicator: an empirical analysis. *Scientometrics, 87*(3), 467-481.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011b). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics, 5*(1), 37-47.

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*(3), 851–872.

Wang, X., & Zhang, Z. H. (2020). Improving the reliability of short-term citation impact indicators by taking into account the correlation between short- and long-term citation impact. *Journal of Informetrics, 14*(2), 101019.

Wang, X., & Zhang, Z. H. (2021). TWCNCI: A new weighted field normalization indicator considering the reliable degree of citation time window. (submitted for publication) (in Chinese)

Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology, 59*(11), 1856-1860.

Comparing Paper Level Classifications across Different Methods and Systems: An Investigation on *Nature* Publications

Lin Zhang[1], Beibei Sun[2], Fei Shu[3] and Ying Huang[4,*]

[1] *linzhang1117@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan 430072 (China)
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan 430072 (China)

[2] *betty_sun@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan 430072 (China)
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan 430072 (China)

[3] *fei.shu@hdu.edu.cn.at*
Chinese Academy of Science and Education Evaluation, Hangzhou Dianzi University, Hangzhou 310018 (China)

[4] *ying.huang@kuleuven.be*
Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven B-3000 (Belgium)
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan 430072 (China)

Abstract

The classification of scientific literature into appropriate disciplines is one of the essential preconditions of valid scientometric analysis and is significant in the practice of research assessment. In this *research-in-progress* paper, we compared the disciplinary classifications of publications in *Nature* based on three different approaches across three different systems: WoS categories provided by Incites based on the disciplinary affiliation of the majority of its references; FOR major and minor fields provided by Dimensions based on the machine learning techniques; and *Nature* subjects constructed by Springer Nature based on the author-selected subject area tags in the journal's tagging system. The preliminary results show that, the single category assignment in Incites based on the disciplinary affiliation of cited references seems quite unreasonable for a large proportion of papers. The comparison results between Dimensions and *Nature* also show significant differences: only 1322 (16%) publications share the same classification results. Further investigations including a closer look at the "inconsistent" results, a deeper insight and methodological analysis is needed in the completion of the study.

Introduction

The classification of science into a disciplinary structure is as old as science itself (Glänzel & Schubert 2003). Classification systems assign scholarly literature into disciplines, which can help us delineate fields and identify their thematic structures(Young & Belanger 1983), and of great significance on scientific evaluations (Shu et al. 2020).

In the last decades, multiple methods have been developed to identify academic disciplines to which publications belong. For an overview of the history of classification systems, one can refer to Gläser et al. (2017). Most classification systems of publications are based on journal assignment, and it is most commonly adopted by major bibliographic databases such as Clarivate Analytics' Web of Science (WoS) and Elsevier's Scopus. However, using such classification systems, all papers published in a given journal are classified in the same discipline (or set of disciplines), which has been treated as a limitation. As stated by Leydesdorff & Bornmann (2016) and Shu et al. (2019), these journal level categories do not provide sufficient analytical clarity to carry bibliometric normalization in evaluation practices.

There are three main solutions to the classification problem on article level. The first is the citation-based classification approach. Waltman & van Eck (2012) clustered almost 10 million publications to 20 research areas based on their citation relations, and the algorithms are continuously optimized (Traag et al. 2019). Klavans & Boyack (2017) compared the accuracies of topic-level taxonomies based on the clustering of documents using direct citation, bibliographic coupling, and co-citation. The second is the text-based algorithmic approach.

Boyack et al. (2011) clustered over two million biomedical publications based on nine different document-document similarity matrices generated from information extracted from their titles, abstracts and subject headings. Eykens et al. (2019) attempted to classify sociology publications with supervised machine learning algorithms based on the textual information from titles and abstracts. The third is the author-selected-based approach. Chinese library classification (CLC) system has been used at the paper-level by asking authors to provide the CLC code when submitting their manuscript (Shu et al. 2019). McGillivray & Astell (2019) also introduced the subject tagging system constructed Springer Nature based on author-selected subject area tags, which is one of the data sources in this study.

In this study, we attempt to assess how citation-based classification (in this case the Incites WoS categories) and text-based algorithmic classification (in this case the Dimensions Fields of Research, FOR) perform on the individual article level, in comparisons with Springer Nature's subject classification based on author-selected subject tagging.

Data and methods

In this exploratory study, we chose research articles and letters[1] published during 2010-2019 in the multi-disciplinary journal *Nature* as the data sample. The three different classification systems and corresponding methods used in this study are briefly introduced as follows:

1) *Web of Science (WoS) Categories*

The WoS schema comprises 250+ categories, and each journal and book covered by the WoS core collection is assigned to at least one WoS category. Additionally, all articles in theWoS categories: '*Multidisciplinary Sciences*', '*Medicine, General and Internal*' are reclassified into specific categories based on the disciplinary affiliation of its cited papers. The results can be obtained from Incites™. After gathering all cited references, along with the respective WoS categories assigned to the journals in which the cited references occur, the target paper is then reclassified to the most frequently occurring category from this distribution. Finally, one publication is only assigned to one WoS category.

2) *Fields of Research (FOR) classification*

The Fields of Research (FOR) classification is a component of the Australian and New Zealand Standard Research Classification (ANZSRC) system developed in 2008. It contains 22 divisions (broad subject area, FOR2) and 159 groups (detailed subsets of these categories, FOR4). FOR categories are built in Dimensions using emulations of the categorisation systems based on machine learning guided by topic experts. In Dimensions, one publication may be assigned to multiple FOR4 or FOR2 categories with the classifying algorithm. The assignments of FOR classifications to each publication in *Nature* can be obtained from the website of Dimensions.

3) *Nature subjects*

Springer Nature constructed a standardized structure for the subject tagging of its own publications (McGillivray & Astell 2019). The subject tagging system is based on author-selected subject area tags, which are chosen at the point of submission and inserted into the article's XML at publication. There are thousands of subject area tags in *Nature*'s system, and they are aggregated to 95 second-level subjects and 8 top-level subject areas. Publications can have as many subject area tags associated with them as the article's authors choose. This information is available on the article online-page in the website of *Nature*, and all subjects linked to each document are listed in an integrated website[2].

For the comparability across different classification systems, a uniform classification system is needed. Since the mapping correspondence between OECD Fields of Research (FOS) and WoS

[1] The document types, Article and Letter, are defined according to the *Nature*'s website. They are both peer reviewed research papers published in *Nature*.

[2] https://www.nature.com/nature/browse-subjects

categories can be obtained from Incites-Help[3], and the correspondence table between OECD FOS and FoR can be obtained from the Australian Bureau of Statistics (ABS) official website[4], OECD classification was chosen as the uniform classification system here. As for the correspondence between *Nature* subjects and OECD, we made a manual match through the following steps: 1) Direct mapping of *Nature* subjects and OECD sub-fields with specific feature words (for instances, *chemistry*, *materials science*, etc.). 2) For those subjects with unclear feature words (for instances, *risk factors*, *biotechnology*, etc.), we manually mapped them to OECD sub-fields by referring to the *Nature*'s top-level areas and viewing related papers. 3) For the three *Nature* subjects with broad coverage (*Engineering*, *Agriculture,* and *Social sciences*), we further crawled each paper's subject area tags, and assign each paper to the related OECD sub-fields according to their subject area tags. Specific data acquisition and mapping methods are outlined in figure 1.

After combining all data obtained from the above three platforms/databases, we finally got the total data sample with 8232 papers published in *Nature* during 2010-2019.

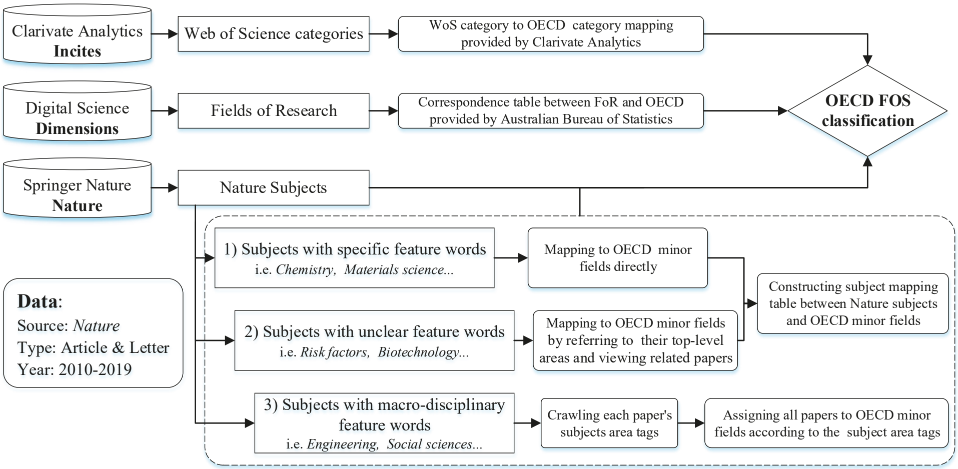

Clarivate Analytics
Incites
Web of Science categories
WoS category to OECD category mapping provided by Clarivate Analytics
Digital Science
Dimensions
Fields of Research
Correspondence table between FoR and OECD provided by Australian Bureau of Statistics
OECD FOS classification
Springer Nature
Nature
Nature Subjects
Data:
Source: Nature
Type: Article & Letter
Year: 2010-2019
1) Subjects with specific feature words
i.e. Chemistry, Materials science...
Mapping to OECD minor fields directly
2) Subjects with unclear feature words
i.e. Risk factors, Biotechnology...
Mapping to OECD minor fields by referring to their top-level areas and viewing related papers
Constructing subject mapping table between Nature subjects and OECD minor fields
3) Subjects with macro-disciplinary feature words
i.e. Engineering, Social sciences...
Crawling each paper's subjects area tags
Assigning all papers to OECD minor fields according to the subject area tags

Figure 1. Outline of the data acquisition and mapping methods

Results

The comparison between Incites and Nature

Firstly, we divided all publications into three groups according to the comparison results based on OECD "subfields". Note that in Incites, each paper is assigned to a unique subfield, while *Nature* allows multi-assignments. The results are presented in figure 2.

Group 1 (Identical): The OECD subfield mapped from Incites and *Nature* is identical. 2309 publications, accounting for 28% of all publications belong to this group.

Group 2 (Partial identical): 4869 (59%) publications are assigned to more than one OECD subfield in *Nature* system, while the subfield mapped from Incites is only one of the multi-assignments. For example, paper A is classified into subfields *a* and *b* in *Nature* system, and the same paper is assigned to only subfield *b* according to Incites.

Group 3 (Inconsistent): 1054 publications (13%) have completely different classification results of the OECD subfield(s) between Incites and *Nature*. For example, paper A is classified into

[3] https://incites.help.clarivate.com/Content/Research-Areas/oecd-category-schema.htm
[4] https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1297.02008?OpenDocument

subfields *a (or* and *b)* in *Nature* system, and the same paper is assigned to subfield *c* according to Incites.

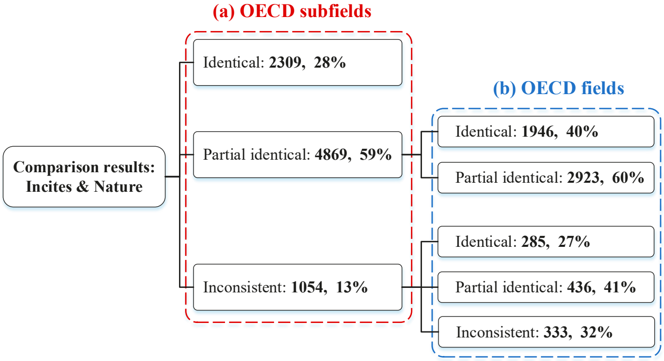


Figure 2. The comparison results between Incites and *Nature* based on mapped OECD (sub)fields

As a second step, we extended the similar comparison based on OECD "major fields" for publications in group 2 and group 3. For publications in group 2, 1946 papers (40%) are with the same OECD major field in both Incites and Nature systems. The rest of publications in this group are assigned to at least two OECD major fields in Nature, and the field assignment from Incites is only one of them. For publications in group 3, 285 papers (27%) are assigned to the same major field across the two systems, and 333 publications (32%) have completely different classifications even when adopting the OECD major fields.

We roughly calculated the proportion of the number of references with the largest subject category to the total number of referenced subject categories in each paper. The proportions of 4217 publications (accounting for 51% of the total sample) are less than 30%, and that of 8076 (98% of the total sample) publications are smaller than 50%, which shows clear unreasonableness of the single category assignment in Incites. Representative examples of publications in group 2 and group 3 have been found to demonstrate the problems of Incites reclassification, but restricted to the length of RIP paper, an in-depth analysis will be provided in the full version. In addition, the categories of cited references are via journals classification. By comparing the journal- and paper- level classifications for the same set of papers and journals, Shu et al. (2019) reported that half of papers could be misclassified in journal classification systems.

The comparison between Dimensions and Nature

All publications are firstly divided into three groups according to the comparison results based on OECD "subfields". Both Dimensions and *Nature* allow multi-assignments, which add to the complexity of comparison. The results are presented in figure 3.

Group 1 (Identical): The OECD subfield(s) mapped from Dimensions and *Nature* are identical. 1322 publications, accounting for 16% of the total sample belong to this group, of which 86% are assigned to one single subfield in both classifications.

Group 2 (Partial identical): This group comprises three sub-groups of publications: 1) The OECD subfield(s) mapped from Dimensions is a subset of subfields mapped from *Nature* (2407 papers, 29% of the total sample). 2) The OECD subfield(s) mapped from *Nature* is a subset of subfields mapped from Dimensions (1350 papers, 16%). 3) The OECD subfield(s) mapped from the two systems are partly identical, but none could be identified as subset of the other (2518 papers, 31%).

Group 3 (Inconsistent): The OECD subfield(s) mapped from Dimensions are completely different with the subfield(s) mapped from *Nature* (635 papers, 8%).

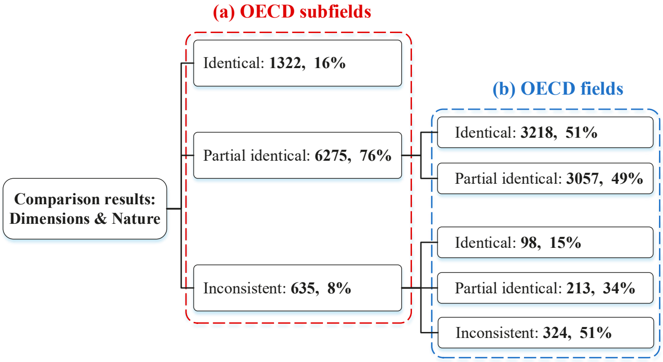


Figure 3. The comparison results of Dimensions and *Nature* based on mapped OECD (sub)fields

We further extended the comparison based on OECD major fields for publications in group 2 and group 3. For publications in group 2, 3218 papers (51%) are with the same OECD major field(s) in both Dimensions and *Nature* systems. 3057 (49%) papers are still partial identical, which can be further divided into: 1) The OECD field(s) mapped from Dimensions is a subset of fields mapped from *Nature* (2405 papers, 79% of the total sample). 2) The OECD field(s) mapped from *Nature* is a subset of fields mapped from Dimensions (565 papers, 18%). 3) The OECD field(s) mapped from the two systems are partly identical, but none could be identified as subset of the other (87 papers, 3%). For publications in group 3, 98 papers (15%) are assigned to the same major field across the two systems, and 324 publications (51%) have completely different classifications even when adopting the OECD major fields. An in-depth analysis of representative publications in group 2 and group 3 will be presented in the full version.

Discussion and conclusion

In this paper, we obtained the disciplinary classifications of publications in *Nature* based on three different approaches across three different systems. The comparison results of Incites and *Nature* show that, the OECD subfield of 2309 (28%) publications mapped from Incites and *Nature* are identical. 4869 (59%) publications are assigned to more than one OECD subfield in *Nature*, while the subfield mapped from Incites is only one of them. And 1054 (13%) publications have completely different classification results of the OECD subfield(s) between Incites and *Nature*. On the one hand, the single category assignment based on the disciplinary affiliation of cited references seems quite arbitrary. On the other hand, the categories of cited references are based on journal-focused solutions. The accuracy has been problematized since the existence of journals that publish papers from multiple research areas and papers published in those journals beyond their fields (Shu et al. 2019; Milojević 2020). Being aware of the characteristics and limitations of the existing categorization of research publications, a bottom-up approach demonstrated in InCites™ Citation Topics was also introduced (Szomszor et al. 2021). Further investigation on this new citation-based, dynamic classification scheme will be performed in future work.

The comparison between Dimensions and *Nature* shows that, there are 1322 (16%) publications whose OECD subfield(s) mapped from Dimensions and *Nature* are identical. 6275 (36%) publications are partial identical. And 635 papers (8%) are completely different. As stated above, the FOR categories are built in Dimensions using emulations of the categorization systems led by supervised machine learning. However, different techniques and methodologies employed in text documents classification have their own limitations. Firstly, the classifier training should be performed on the basis of large number of training text terms, which is very laborious. And if the predefined categories changed, these methods must collect a new set of training text terms. Secondly, most traditional methods haven't considered the semantic relations between words, so the accuracy of related classification methods still remain to be improved. In addition, although it is assumed that the classification of Springer Nature is more accurate than the other two classification systems considering the authors' self-selected subject area tags, the accuracy still needs to be further evaluated.

In this preliminary study, we briefly demonstrate the comparison results of paper-level classifications across Incites, Dimensions and *Nature* subjects. Further investigations including a closer look at the "inconsistent" results, a deeper insight and methodological analysis is needed in the completion of the study. Our final results will be presented at ISSI 2021 if we are given the possibility to do so.

Acknowledgement

The authors would like to acknowledge support from the National Natural Science Foundation of China (Grant Nos. 71573085, 71974150, 72004169), the National Laboratory Center for Library and Information Science in Wuhan University, and the project "Interdisciplinarity & Impact" (2019-2023) funded by the Flemish Government.

References

Boyack, K. W., Newman, D., Duhon, R. J., et al. (2011). Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. *Plos One, 6*(3): e18029.

Eykens, J., Guns, R., & Engels, T. C. E. (2019). Article Level Classification of Publications in Sociology: An Experimental Assessment of Supervised Machine Learning Approaches. In *17th International Conference on Scientometrics & Informetrics.* Rome, Italy. pp: 738-743.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56*(3): 357-367.

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics, 111*(2): 981-998.

Klavans, R., & Boyack, K. W. (2017). Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge? *Journal of the Association for Information Science and Technology, 68*(4): 984-998.

Leydesdorff, L., & Bornmann, L. (2016). The operationalization of "fields" as WoS subject categories (WCs) in evaluative bibliometrics: The cases of "library and information science" and "science & technology studies". *Journal of the Association for Information Science and Technology, 67*(3): 707-714.

McGillivray, B., & Astell, M. (2019). The relationship between usage and citations in an open access mega-journal. *Scientometrics, 121*(2): 817-838.

Milojević, S. (2020). Practical method to reclassify Web of Science articles into unique subject categories and broad disciplines. *Quantitative Science Studies, 1*(1): 183-206.

Shu, F., Julien, C.-A., Zhang, L., et al. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics, 13*(1): 202-225.

Shu, F., Ma, Y., Qiu, J., et al. (2020). Classifications of science and their effects on bibliometric evaluations. *Scientometrics, 125*(3): 2727-2744.

Szomszor, M., Adams, J., Pendlebury, D. A., et al. (2021). Data categorization: understanding choices and outcomes.

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep, 9*(1): 5233.

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12): 2378-2392.

Young, H., & Belanger, T. (1983). *The ALA Glossary of Library and Information Science*: Chicago: American Library Association.

The prevalence and impact of different types of open access articles from China and USA

Lin Zhang[1], Yahui Wei[2], Ying Huang[3,*] and Gunnar Sivertsen[4]

[1] *linzhang1117@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan (China)
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)

[2] *wei_ya_hui@163.com*
School of Information Management, Wuhan University, Wuhan (China)
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)

[3] *ying.huang@kuleuven.be*
Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven (Belgium)
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)

[4] *gunnar.sivertsen@nifu.no*

Nordic Institute for Studies in Innovation, Research and Education, Oslo (Norway)

Abstract

Compared to the traditional subscription-based business model, Open Access (OA) publishing has many advantages. It can make science accessible to a wider public and expand individual researchers' influence and increase the potential impact of their research. This study covers all articles and reviews published from China and the USA from 2010 to 2019 with the aim of analyzing the prevalence and impact of different types of OA research articles. We find that the USA has a higher percentage of OA publications than China. American OA articles are more frequently in the Green and Bronze types of OA, while China publishes relatively more in Gold journals covered in DOAJ. Regarding citations, hybrid OA publications from American researchers have the highest impact, while the non-OA publications articles have the lowest impact. Green OA publications from Chinese researchers have the highest impact, while those appearing in Gold DOAJ journals have the lowest impact. The differences can be explained by national research evaluation and OA policies.

Introduction

Open Access (OA) is a policy for change in the international infrastructure of scientific publishing to make publications freely available on the public internet. The change was initiated from university libraries and taken on by research policy at the governmental level. The policy is now partly supported by the international academic communities and fully by the major scientific publishers, who are turning their business model from subscription fees to author payment.

The focus of our study is on the OA development in the two largest contributing countries to the world of science. With its fast and continuous economic and scientific development, China has become a prominent and influential nation in science (Zhou et al., 2008). Since 2016, the number of researchers and scholarly articles in China have surpassed the corresponding numbers of the USA (Tollefson, 2018). As the two largest countries by the number of scientific publications, the policies and developments in the USA and China towards OA play an important role in the change of the scientific publishing infrastructure internationally.

Both American and Chinese authorities have been continuously active in the promotion of OA. In 2013, president Barack Obama's administration introduced a policy that requires taxpayer-funded research to be made freely available online within 12 months of its publication in a journal (Subbaraman, 2019). The National Institutes of Health (NIH) in the USA has also allowed a maximum of 12-month embargo to support Green OA (National Institutes of Health, 2015). In 2014, the National Natural Science Foundation of China (NSFC) and the Chinese Academy of Sciences (CAS) issued a statement announcing that their sponsored researchers should deposit their articles into online repositories and make these articles publicly accessible within 12 months of publication (Van Noorden,2014). At the 14th world OA conference in

Berlin in 2018, China's National Science Library (NSL), National Science and Technology Library (NSTL), and NSFC announced their support for Plan S (Schiermeier, 2018). During the meeting, Xiaolin Zhang, chair of the Strategic Planning Committee of the NSTL, said that the NSFC, NSTL and NSL will all support the government's request to make research papers open immediately after publishing. They also announced their support for a wider range of flexible and inclusive measures to achieve this goal, which means that China could adopt other OA types than Gold OA to achieve Plan S. At the same time, the USA federal agencies stayed with policies developed in 2013. "*We don't anticipate making any changes to our model*," said Brian Hitson of the USA Department of Energy in Oak Ridge, Tennessee, who directs this agency's public access policy (Rabesandratana, 2019). The statement indicated that the USA still prefers the publishing model of Green OA. However, the Scholarly Publishing and Academic Resources Coalition, a group that represents more than 200 USA college and university libraries, said that its members "*wholeheartedly endorse updating current policy and eliminating the unnecessary 12-month waiting period* (Subbaraman, 2019)."

There are some similarities in the OA policies of the two countries. For example, both require researchers to deposit their funded articles online for free within 12 months after publication. But in recent years, some institutions are beginning to adjust their policies to achieve OA as soon as possible after the articles are published.

A comparison has not been made so far of the OA output of the two major publishing countries and the impact of different types of OA publications in relation to national policies. We investigate the prevalence and impact of different types of OA publications from the two countries to answer these questions:

- Q1: What is the prevalence of OA publishing in China[1] and the USA?
- Q2: What types of OA publishing are preferred in the two countries?
- Q3: What is the citation impact of the different types of OA publications in the two countries?

Data and methods

Types of OA

This study is based on information made available in Web of Science by *Clarivate* in collaboration with *Our Research* and their Unpaywall database.

This source of information distinguishes between five types of OA articles according to the license policy of the journals:

1) *DOAJ Gold*. Articles published in journals listed on the Directory of Open Access Journals (DOAJ). To be listed on the DOAJ, all the articles in these journals must have a license in accordance with the Budapest Open Access Initiative.
2) *Hybrid (Other Gold)*. Hybrid open access articles are those identified as having a Creative Commons (CC) license by the Unpaywall Database but are not in journals listed on the DOAJ. Hybrid journals are based on subscriptions but give free access to individual articles against author payment.
3) *Bronze*. The license for these articles is either unclear or is in the Unpaywall Database as articles without CC license that are free-to-read on the publisher's site.
4) *Green*. including Green Published, the final published versions of articles hosted in an institutional or subject-based repository, and Green Accepted, accepted final versions of manuscripts hosted on a repository without the publisher's copyediting or typesetting.
5) *Non-OA*. All other research articles.

[1] China in this paper refers to mainland China.

Data collection and processing

Data for this study were collected from InCites, an analytical tool based on Web of Science[2]. To cover all different types of OA research articles, we selected all articles and reviews published by at least one Chinese author (regardless of the author's position) and all articles and reviews published by at least one American author from 2010 to 2019. The 22 disciplines of the Essential Science Indicators (ESI) were selected for the subject classification. Since InCites does not provide information about the first author/corresponding author and whether there is international collaboration in each document, we added this information by downloading all documents from Web of Science. The data processing procedures are shown in Figure 1.

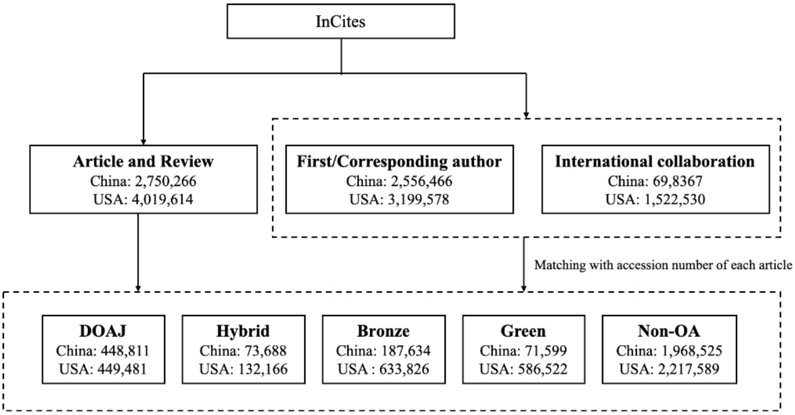


Figure 1. The procedures of data processing.

When multiple open access versions of an article are available, Web of Science prioritizes publisher-hosted content (i.e., gold, hybrid, or bronze), then the most complete version (i.e., for green OA, published version over accepted version over submitted version) (Bosman & Kramer, 2018). Web of Science only includes the 'best' OA location as determined by this algorithm. However, many articles are still classified into multiple OA types, which may affect the analysis. To ensure the accuracy of our study, we relabeled the OA types of all multiple assignment articles to ensure that each article belongs to only one type of OA. Following the example of Piwowar (2019), if an article occurs in both a Gold journal and an OA repository, it is classified as Gold, not Green. Adding other principles from Web of Science[3], the instances of multi-assignment are treated this way in our study: 1) DOAJ in combinations with Green is classified as *DOAJ*; 2) Hybrid in combinations with Bronze or Green are classified as *Hybrid*; 3) Bronze in combinations with Green is classified as *Bronze*.

Results

The prevalence of OA in China and the USA

Figure 2 shows the total number of research articles, the number of OA research articles, and the percentage share of these OA research articles in China and the USA from 2010 to 2019. The OA gap between China and the USA, both in relative and absolute numbers, has gradually decreased and the policy-driven transition to OA is evident for both countries, at least until 2017, after which it stagnated in China and decreased in the USA. This remarkable change will

[2] https://incites.clarivate.com/#/landing. InCites dataset updated Dec 17, 2020. Includes Web of Science content indexed through Nov 30, 2020.
[3] https://clarivate.com/webofsciencegroup/solutions/open-access/

be discussed and analyzed in the following as we have a closer look at the different types of OA research articles.

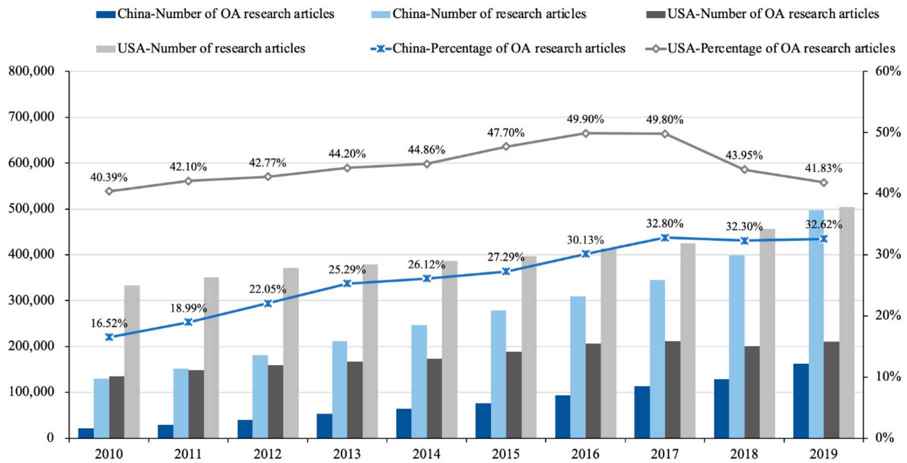
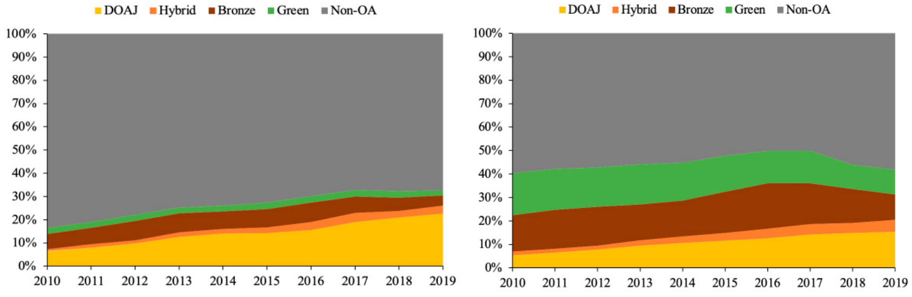


Figure 2. Total number of research articles, number of OA research articles, and the percentage of OA research articles from 2010 to 2019.

The percentage shares among five types of OA research articles in China and the USA from 2010 to 2019 is shown in figure 3. After 2017, the percentage of Green and Bronze OA research articles in the USA began to decrease, while the percentage of DOAJ and Hybrid OA research articles continued to increase. The declining trend may be partly related to the time lag of different types of OA research articles. There is no lag of DOAJ and Hybrid OA because the articles become OA immediately after publication. However, for Green and Bronze OA, the lag is explainable. Authors often self-archive (upload their paper to a repository) months or years after the official publication date of the articles, typically because many journals have policies that authors must wait a certain length of time (the “embargo period”) before self-archiving (Piwowar et al., 2019).

China and the USA, although with similar OA policies, have quite different orientations towards the types of OA research articles. China has the largest percentage of OA research articles in DOAJ journals, while the USA has the largest percentage of OA research articles in Green and Bronze. Some explanations for these differences are available. Firstly, OA policies in China are mostly advisory. Researchers are encouraged to store their publications in the institutional or subject-based repository. There is no mandatory time requirement. The articles can be uploaded at any time. In contrast, some OA policies in the USA are mandatory. For example, the National Institutes of Health Public Access Policy has developed from unofficial regulations to mandatory OA policies to ensure the policy implementation’s legal effect[4]. A second explanation is the impact of research evaluation policy in China. For the last decade, China’s research evaluation and funding policies have had a strong focus on quantitative indicators with incentives to publish in journals covered by the Web of Science (Zhang & Sivertsen, 2020; Quan et al., 2017). Simultaneously, some DOAJ or Hybrid OA journals indexed in WoS (often with high article processing charges) have expanded their annual volume of articles and lowered the threshold for publication, thus affecting Chinese researchers’ choice to a certain extent. The rapidly increasing number of journals in DOAJ (from 300 in 2003 to 15,000 presently) can explain the increasing role of this OA type in both countries.

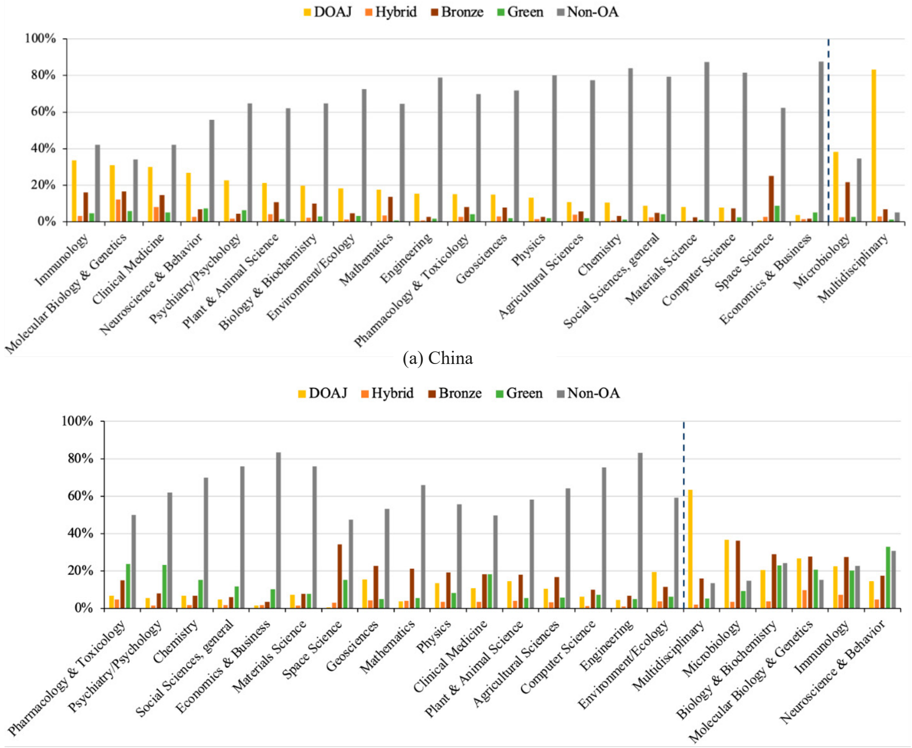
[4] National Institutes of Health, "Request for Information: NIH Public Access Policy", available at https://publicaccess.nih.gov/comments.htm.



(a) China (b) USA

Figure 3. Percentage of five types of OA research articles from 2010 to 2019.

Figure 4 illustrates the percentage of the different OA types of research articles in the 22 ESI disciplines from 2010 to 2019. Most disciplines mainly publish research articles in the form of Non-OA. Multidisciplinary Sciences (representing e.g. the large DOAJ journal *PLoS One*) and Microbiology have the highest percentage of DOAJ research articles both in China and the USA. Among the OA research articles, DOAJ is the most prevalent type among Chinese researchers in 20 disciplines, while Bronze and Green are more prevalent OA types among American researchers in 20 disciplines.



(a) China

(b) USA

Figure 4. The percentage of the five OA types of articles in different disciplines from 2010 to 2019.

OA research articles with authors from China and the USA in the lead

Generally, the first author or corresponding author can be regarded as having the main responsibility and contribution to an article. We will now consider those articles only where the first author or corresponding author is from China or the USA. Figure 5 shows percentage of the articles within the five types of OA from a country where the authors in lead are from the same country. With this measurement, we can indicate the more typical OA practices of each of the countries. China has more distinctive OA practices towards Non-OA (94.84%) and DOAJ (92.95%). Green OA is less distinctive for China because authors from other countries tend to be in the lead in these articles (see below). Contrary to China, Green OA research articles are distinctive for the USA, while DOAJ and Hybrid OA are not. Our results are consistent with the results of a large-scale questionnaire survey by Professor Carol Tenopir's team (2017). They investigated researchers' attitudes and behaviors towards gold OA in the USA. The results showed that most respondents hold neutral to somewhat negative opinions towards gold OA. They believed that articles published in OA journals are of lower quality than those published in subscription-based journals.

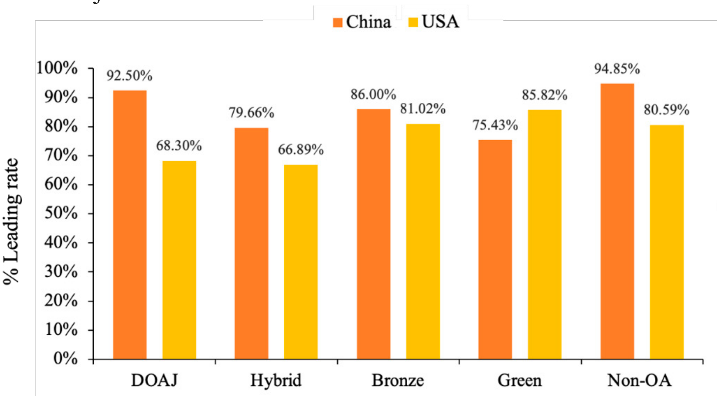
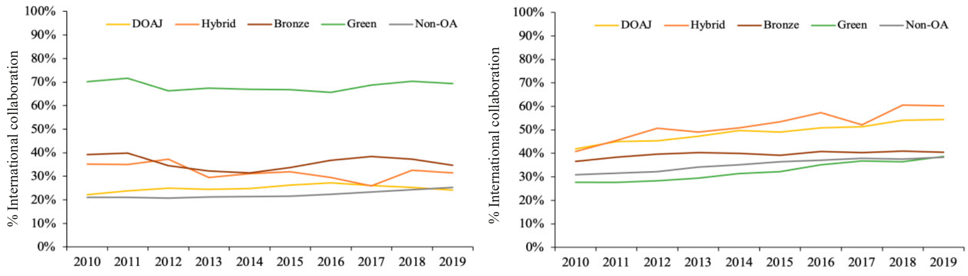


Figure 5. Percentage of the five types of OA research articles led by China and the USA.

Note: Percentage calculation method: Percentage of the DOAJ research articles led by China= Number of DOAJ research articles with Chinese researchers as the first or corresponding author /Total number of DOAJ research articles published by Chinese researchers * 100%.

In articles with *international collaboration*, the choice of OA type and journal will need to be agreed by co-authors from different countries. These articles may therefore less distinctly represent the OA practices of the USA and China, in contrast to the articles analysed in Figure 5. Figure 6 shows the percentage distribution of articles with *international collaboration* among the five OA types from 2010 to 2019. The proportion of articles with Green OA from China is now highest, far higher than the other four types of articles, indicating that the Green OA practice is mainly due to collaboration with scholars from other countries. The percentage of Non-OA research articles from China is low. This is consistent with the distinctive traits presented above. Non-OA and DOAJ OA are distinctive Chinese practices. The results also confirm that hybrid OA is not a distinctively American practice while Green OA certainly is.



(a) China (b) USA

Figure 6. Proportion of OA types in articles with international collaboration from 2010 to 2019.

The citation impact of different types of OA research articles in China and the USA

We have selected four citation impact indicators from InCites to compare the five types of OA research articles and the two countries:

- *Average Citation*. The average number of citations per paper.
- *%Documents in Top1%*. Percentage of publications in the top 1% most highly cited publications in the world, based on citations by category, year, and document type.
- *Category Normalized Citation Impact (CNCI)*. Citation impact (citations per paper) normalized for the subject, year, and document type.
- *Journal Normalized Citation Impact (JNCI)*. Citation impact (citations per paper) normalized for the journal, year, and document type.

Figure 7 shows the four impact indicators among the five types of OA research articles. Figure 7(a) indicates that the average citation impact within all five types of research articles is higher for the USA than for China. Green OA research articles have the highest impact in both countries followed by Bronze OA articles. These types of OA can be related to traditional subscription-based journals. The same is true for Hybrid OA. The higher impact of these categories compared to Non-OA, is therefore remarkable, particularly for the USA. OA seems to provide more impact. The impact of DOAJ OA research articles is relatively low for both countries, particularly for China where this type of OA is a distinctive and abundant practice.
Moving from average citation impact to the field-normalized indicators in Figure 7 (b) (extremely high impact articles) and Figure 7 (c) (all articles), the picture slightly changes. By these indicators, Hybrid OA research articles score the highest among articles from the USA while Green OA research articles score the highest among articles from China. We already know from the results above that these OA practices are the less distinctive types for both countries and that they are most influenced by international collaboration. Articles with international collaboration are in general more cited. The largest differences in citation impact between the two countries are now found in the Hybrid and Bronze types while there is almost no difference in the Green and Non-OA types. This is a remarkable finding showing the need for further investigation of APC costs related to OA in the publishing practices of different countries. We will follow up in another study.
The JNCI indicator is a similar indicator to the CNCI indicator used above, but instead of normalizing per subject area or field, it normalizes by the average citation rate of articles in the journal in which the article was published. This indicator is valuable for comparing the two countries within journals with the same OA policy. Figure 7 (d) shows that this indicator tends to level out the differences between the two countries in citation impact in all OA types except Hybrid. Again, the least distinctive OA type depending the most on international collaboration (Hybrid for the USA, Green for China) is the most favourable when comparing the two countries. More generally, on the background of the results of the three other citation indicators

presented above, this last indicator indirectly shows that the two countries tend to publish in *different* journals within the *same type* of OA. As an example, we observe that the differences in citation impact between the two countries within DOAJ in Figure 7a-c diminishes in Figure 7d where citations are normalized relative to the articles in the same journal. It seems that the articles from China in DOAJ are published in journals with lower impact.

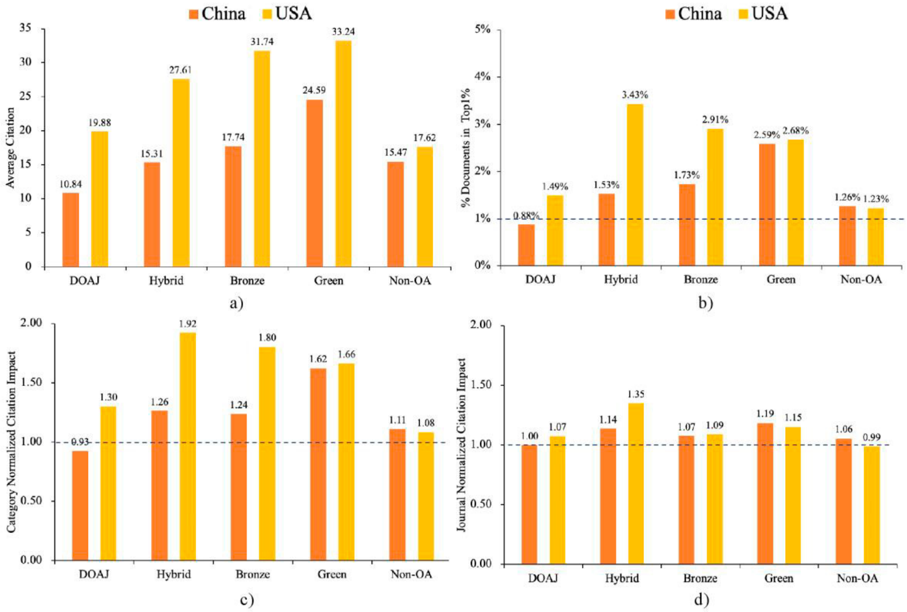
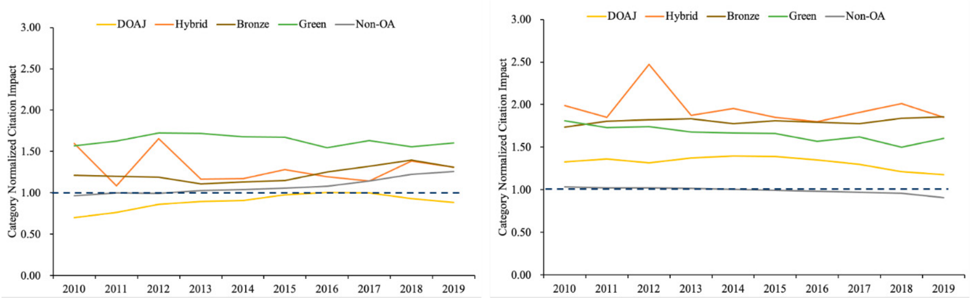


Figure 7. Impact indicators of the five types of OA research articles.

The CNCI indicator is frequently used in comparisons of countries with different research profiles. We chose this indicator as representative in a further analysis of trends during the period studied. Figure 8 displays the CNCI of the five types of OA research articles from both countries in the period from 2010 to 2019. The general differences between the two countries according to this indicator were already observed above. Here, we observe some interesting similarities and differences in the trends. The major difference is that the CNCI of Non-OA research articles increases for China while being stable or decreasing for the USA. China is even surpassing the impact of USA in these traditional journals. There is also an increase for China in the Bronze type of articles, which also may appear in traditional journals. This might indicate that China is not gaining from the so-called ‘citation advantage’ of OA. However, the impact of China in the Green type is high for China and stable over time while decreasing for the USA. The explanation these differences might that different journals are chosen within each type (as seen in Figure 7d). It might be that American articles are published in less influential journals whenever the Non-OA alternative is chosen. There is nonetheless one clear similarity. The impact of DOAJ OA research articles is low for both countries and falling towards the end of the period, probably because of the rapid growth of new DOAJ journals.

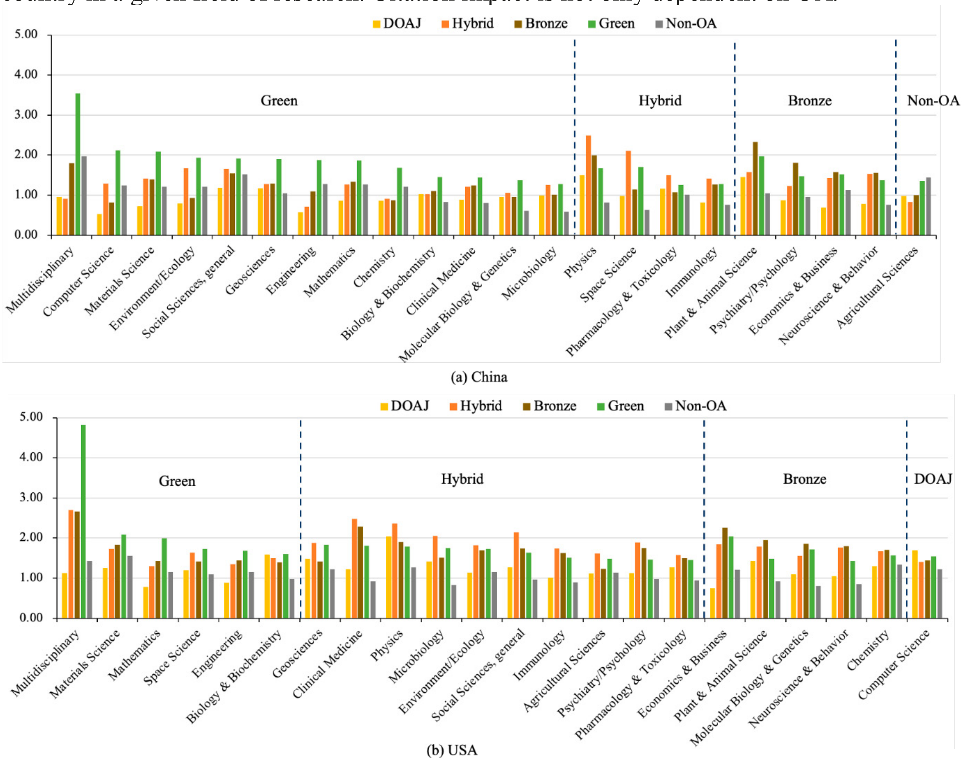


(a) China

(b) USA

Figure 8. The field-normalized citation impact (CNCI) of articles with different types of OA from 2010 to 2019.

The impact of different types of OA research articles in different disciplines is also worthy of attention. Figure 9 shows the CNCI of different types of OA research articles in the ESI 22 disciplines. To aid the interpretation of the results, we divide the 22 disciplines into four categories according to the most prevailing OA type both in China and the USA. The specific subject classification can be viewed from Figure 9. The most interesting observation here is that as many as half of the disciplines occur in different OA types by country when measured according to the citation impact of the country. The results indicate that citation impact from OA is not only contingent on the type of OA practice, but also on the relative impact of the country in a given field of research. Citation impact is not only dependent on OA.



(a) China

(b) USA

Figure 9. The CNCI of different types of OA research articles in 22 disciplines.

Discussion and conclusion

Our results show that policy intervention has an important impact on the development of OA. Firstly, we find that the percentage of OA research articles in the USA is higher than in China. The explanation may be that China's OA policies are mostly advisory, not mandatory. Some studies have showed that most Chinese researchers approve of OA in principle (Ren, 2015; Wang, 2013), but in practice, they are reluctant to submit papers to OA journals or to institutional repositories (Yuan & Zhang, 2016). As stated by Xu et al. (2016), many Chinese researchers are also sceptical and confused about OA publishing. However, we know from our finding in Figure 2 that this situation must be changing. A more recent questionnaire-based survey from the Chinese Academy of Science and Technology for Development (CASTED) （2020) showed that 65% of the Chinese researchers expressed their willingness to publish papers in OA journals.

Both of the governments of China and the USA promote Green OA, but the relative prevalence of the types of OA in the two countries is different. Green is the most prevalent OA type in the USA, while DOAJ is the most prevalent OA type in China. Zhang Xiaolin, chairman of the Strategic Planning Committee of the Chinese National Science and Technology Library, has said that the impression that OA has little influence in China is misleading.[5] Since 2014, funders and research institutions in China have encouraged and funded scientists to publish their papers in open-access formats and archive manuscripts openly online (Schiermeier, 2018). In practice, all research funders in China allow researchers to use research grants for publishing expenses, including article processing charges (APCs) for OA journals (Zhang, 2014). This may have led to an increase in the number of DOAJ articles in China. For example, our recent research found that one of the fastest expanding publishers of DOAJ journals, the Multidisciplinary Digital Publishing Institute (MDPI), publishes the more Chinese DOAJ articles than any other publisher within DOAJ. Since 2010, the proportion of Chinese papers in DOAJ published by MDPI has increased almost exponentially. The MDPI journal *Sustainability* has published more than 30,000 articles since 2016 among which Chinese mainland institutions contributed with about 17,000 articles. However, the expansion of Chinese articles within DOAJ has also created concern in China. As an example, the DOAJ journal *Tumor Biology* retracted 107 papers in one single notice, and most of them were from China (Watch, 2017). Tang (2019) found that China has published 8% of the world's scientific articles, but by 2017 was responsible for 24% of all retractions.

Concerned about these ethical issues, at the beginning of 2020, China issued a policy stipulated that "*For a single paper whose publication expenditure exceeds RMB 20,000, the academic committee of the corresponding author or first author's institutions must review the necessity of the publishing the paper*" and "*Papers published in academic journals on the 'blacklist' and early warning list shall not be included in the special funds for national science and technology plan projects*". This policy shows that China is not only concerned about the costs of article processing charges in Gold OA journals, but also of their consequences for research ethics and quality. Nevertheless, costs may be a concern *per se*: 79.2% of Chinese researchers selected "high publishing cost" as the reason for reluctance to publish papers in OA Journals, according to CASTED's survey (Chinese Academy of Science and Technology for Development, 2020). Recently, the National Science Library of the Chinese Academy of Sciences released the "International Journal Early Warning List (Trial)", which contains 65 journals indexed by Web of Science. Among them, 41 are OA journals and only 7 journals are Non-OA journals[6]. The 65 journals on the list published more than 50,000 Chinese researchers' articles in 2020, accounting for half of articles in the same journals[7]. Among others, *Sustainability* appears on

[5] https://council.science/current/blog/open-access-in-china-interview-with-xiaolin-zhang-of-the-national-science-library/

[6] https://mp.weixin.qq.com/s/xbyJFtR2lezv6CyRrkxsdA

[7] https://baijiahao.baidu.com/s?id=1688277975775326580&wfr=spider&for=pc

the journal list. And according to the China Publish Ecological Report (2020), the number and proportion of articles published by Chinese researchers in *Sustainability* declined for the first time in 2020. Similar lists of journals with warnings are created by and widespread among Chinese institutions. Even the two largest and most famous OA journals in the world, *PLOS ONE* and *Scientific Reports*, are frequently on the Chinese lists of controversial journals.

Our study shows that OA research articles in general have clearly more impact than Non-OA research articles. These results are consistent with previous findings (Wang et al., 2015; Zhang, & Watson, 2017). An explaining factor can be the 'citation advantage' of OA, namely that scholars are much more likely to read and cite easily accessible articles that those that are behind paywalls (Piwowar, 2018). Nevertheless, our study has shown the picture is more complex when comparing different OA types and two countries. Citation impact is not only contingent on OA.

The relatively high impact of Hybrid OA research articles, particularly for the USA, may be related to the characteristics of hybrid journals, which allow the authors to choose whether to make their articles OA to public. A phenomenon called "selection bias postulate" suggests that authors choose to pay APC for only their most potentially impactful work (Craig et al., 2007).

We also found that Green OA research articles published by Chinese researchers have the highest impact, while their DOAJ OA research articles have the lowest impact. There are several possible explanations for this observation. Firstly, the institutional repository is one of the most widely used forms of Green OA. Among the 33 registered repositories in Open DOAR in China, twenty-six of them are from the internationally influential institutes of the Chinese Academy of Sciences (CAS) (Zhong & Jiang, 2017). In our data, 19.2% of the Green OA research articles come from the CAS institutes. The average number of citations and the CNCI of Green OA research articles published by CAS are higher than the average of all research articles in China (CNCI: 1.85 versus 1.1) The influence of the CAS institutes on the indicator may affect the general impact of Green OA articles from China. Secondly, according to the results of the above-mentioned questionnaire survey conducted by the CASTED (2020), 81.8% of the Chinese respondents published OA articles due to "recognition by the evaluation system." It seems that some researchers have published low-quality articles for the needs of evaluation, which may affect the generally low impact of DOAJ OA articles.

However, we observed the relatively lower impact of the DOAJ type OA for both countries, but at clearly different levels. From a Chinese perspective, it is important to know DOAJ OA journals indexed by WoS have been given much weight in the evaluation systems in China until recently (Zhang & Sivertsen, 2020). Hence high quantities of articles with low impact could be expected. The result has been that publishing in DOAJ Gold journals has lost prestige in China although the quantity has risen (Archambault et al., 2013). For USA on the other side, it seems from our results, that by incentivizing or mandating the OA Green alternative, the researchers have been able to remain publishing in the traditionally most impactful journals offering the Hybrid or Bronze alternatives.

Acknowledgement

The authors would like to acknowledge support from the National Natural Science Foundation of China (Grant Nos. 71573085, 71974150, 72004169), the National Laboratory Center for Library and Information Science in Wuhan University.

References

Archambault, É., Amyot, D., Deschamps, P., Nicol, A., F., Rebout, L., & Roberge, G. (2013). Proportion of open access peer-reviewed papers at the European and world levels—2004-2011. *European Commission*,1-24.

Bosman, J., & Kramer, B. (2018). Open access levels: a quantitative exploration using Web of Science and oaDOI data. *Peer J Preprints*, e3520v1.

Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J., & Amin, M. (2007). Do open access articles have greater citation impact? *Journal of Informetrics*, 1(3), 239–248.

Chinese Academy of Science and Technology for Development. (2020). *Cognition, attitude and behavior of Chinese researchers towards open access*.

National Institutes of Health. (2015). National Institutes of Health plan for increasing access to scientific publications and digital scientific data from NIH funded scientific research. Retrieved from https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., & Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, 6, e4375.

Piwowar, H., Priem, J., & Orr, R. (2019). The Future of OA: A large-scale analysis projecting Open Access publication and readership. *BioRxiv*, 795310.

Quan, W., Chen, B. K., & Shu, F. (2017). Publish or impoverish: An investigation of the monetary reward system of science in China (1999–2016). *Aslib Journal of Information Management*, 69(5), 486–502.

Ren, X. (2015). The quandary between communication and certification: individual academics' view on Open Access and open scholarship. *Online Information Review*, 39(5), 628–697.

Rabesandratana, T. (2019). Will the world embrace Plan S, the radical proposal to mandate open access to science papers? *Science*, 363(6422), 11.

Schiermeier, Q. (2018). China backs bold plan to tear down journal paywalls. *Nature*, 564(7735), 171-173.

Subbaraman, N. (2019). Rumours fly about changes to US government open-access policy. *Nature*.

Tenopir, C., Dalton, E. D., Christian, L., Jones, M. K., McCabe, M., Smith, M., & Fish, A. (2017). Imagining a gold open access future: Attitudes, behaviors, and funding scenarios among authors of aca- demic scholarship. *College & Research Libraries*, 78(6), 824–843.

Tollefson, J. (2018). China declare world's largest producer of scientific articles. *Nature*, 553, 390.

Tang, L. (2019). Five ways China must cultivate research integrity. *Nature*, 575, 589-591.

The 2020 China Publish Ecological Report (Humanities and Social Sciences). (2020). Retrieved from https://mp.weixin.qq.com/s/N4ZdND0lCkPNLSCO1bOVZQ.

Van Noorden, R. (2014). Chinese agencies announce open-access policies. *Nature News*.

Wang, P. (2013). A survey on cognition of web-based academic communication behavior by studying on researchers in social science and humanities in China. *Library and Information*, 19(5), 112–118.

Wang, X., Liu, C., Mao, W., & Fang, Z. (2015). The open access advantage considering citation, article usage and social media attention. *Scientometrics*, 103(2), 555-564.

Watch, R. (2017). A new record: Major publisher retracting more than 100 studies from cancer journal over fake peer reviews.

Xu, J., Nicholas, D., Su J.，Zeng, Y. X. (2016). Are Open Aces Journal Trusted by Chinese Scholars. *Geomatics and Information Science of Wuhan University*, (41):131-135.

Yuan, S. B., & Zhang, H. (2016). Self-storage participation behavior of researchers: Qualitative research based on interviews. *Information and Documentation Services*, 13(3), 80–84.

Zhou, P., Thijs, B., & Glänzel, W. (2008). Is China also becoming a giant in social sciences? *Scientometrics*,79(3), 593–621.

Zhang, X. (2014). Development of open access in China: strategies, practices, challenges. *Insights*, 27(1):45-50.

Zhong, J., & Jiang, S. (2016). Institutional repositories in Chinese open access development: Status, progress, and challenges. *The Journal of academic librarianship*, 42(6), 739-744.

Zhang, L., & Watson, E. M. (2017). Measuring the impact of gold and green open access. *Journal of Academic Librarianship*, 43(4), 337–345.

Zhang, L., & Sivertsen, G., 2020. The New Research Assessment Reform in China and Its Implementation. *Scholarly Assessment Reports*, 2(1), p.3.

Productivity patterns, collaboration and scientific careers of authors with retracted publications in clinical medicine

Qin Zhang[1] and Hui-Zhen Fu[2]

[1] *zhangqin20@mails.tsinghua.edu.cn*
Department of Information Resources Management, School of Public Affairs, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou 310058 (China)
School of Public Policy & Management, Tsinghua University, Beijing 100084 (China)

[2] *fuhuizhen@zju.edu.cn*
Department of Information Resources Management, School of Public Affairs, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou 310058 (China)

Abstract

Problematic publications and scientific misconduct have raised global concern, in which the authors have played a key role. Since research misconduct occurs more frequently in the clinical medicine field, we took this field as an example to explore productivity patterns, collaboration and scientific careers of authors at the individual level. The application of Lotka's law (C=0.886, n=3.854) showed a loose collaboration network and a predominant aggregation of multi-retracted authors. Most authors were retracted only once while a small group of authors produced most retracted publications. Authors with most retractions appeared to have stable collaboration with certain individuals. Two typical publication patterns of retractions in different stages of scientific career have been found: (1) committing scientific misconduct in the early career, mainly for promotion, and (2) participating in scientific misconduct when mature in the career, most due to neglecting the duty as a supervisor. The team culture, regulations, publication policies, and joint efforts by the scientific community have a direct influence on the occurrences of scientific misconduct. Conclusions are made with a focus on the need for more actions to prevent scientific misconduct and to strengthen scientific norms and integrity.

Introduction

Scientific research relies on its integrity and scientific misconduct has raised global concerns (Zhang & Grieneisen, 2012). Retracted publications refer to publications that are retracted by the scientific or publication community in concerns of errors, fraud, or other problems in the original publications. Evidence has shown that most of the retracted publications in health, life, and medical sciences were retracted due to research misconduct (Stavale et al., 2019). The rise of retracted publications has raised increasing discussions on scientific misconduct (Mongeon & Larivière, 2016), especially in the medical field (Zhang & Grieneisen, 2012). Repeating fraudulent aspect of funds caused the waste of public resources, but also harmed public trust in scientific research (Sabir, Kumbhare, Parate, Kumar, & Das, 2015). In the field of clinical medicine, flawed and fraudulent researches could lead to inappropriate clinical practice that harmed the safety and health of research participants, ineffective or even harmful treatment to the patients (Gupta, Sahani, Mohan, & Wander, 2013), posing a threat to the public health. What's worse, it may take a great deal of time and effort to discover and retract erroneous or fraudulent publications, while negative impacts of those work could continue by citation in subsequent works without alarming the problems (Budd, Sievert, Schultz, & Scoville, 1999; Hamilton, 2019; Steen, 2011). Therefore, analysis of the nature and dynamics of retracted publications in clinical medicine could contribute to the governance of scientific misconduct and provide profound meaning to the general public.

Previous studies have noticed the recidivism of authors and the unbalanced distribution of retracted publications and in medical field, namely a relatively small number of authors that were observed to have a higher repeating probability of retractions and have contributed to a larger portion of retractions (Kuroki & Ukawa, 2018). Authors with multiple retracted publications are also recognized as “authors with multiple retractions” (Foo, 2011), or “repeat offenders” (Lei & Zhang, 2018). In surgery journals, a quarter of the authors had more than one retraction and it usually happened for the same reasons and even in the same journal (Cassao, Herbella, Schlottmann, & Patti, 2018). Since most authors with multiple retractions were closely associated with scientific misconduct (Mistry, Grey, & Bolland, 2019), it is worth further exploration to analyze their motivation and pattern of producing retracted publications.

Previous studies have mainly revealed the publication and collaboration characteristics of authors with multiple retractions. It has been found that the majority of authors with multiple retractions tended to increase their publications until the first retraction (Mistry, Grey, & Bolland, 2019), and they were also collaboration prone (Tang, Hu, Sui, Yang, & Cao, 2020). Case studies focused on top multi-retracted authors, like Scott Reuben, Joachim Boldt, and Yoshitaka Fujii with the most retractions in clinical medicine, have analyzed the journal responses of their retracted publications (Elia, Wager, & Tramer, 2014; McHugh & Yentis, 2019; Saikia & Thakuria, 2019). Nevertheless, the motivations and factors explaining for individual performance in the problematic publications are remained to be discussed. Furthermore, while co-authors were believed to have a major role in ensuring the integrity of the research and tackling the misconduct (Luther, 2010; Sabir et al., 2015), the impact of collaboration between authors with multiple retractions are also lefted to be explored based on the observed collaboration characteristics.

To give insight of the patterns and motivation of individual scientific misconduct, our study was designed to explored the productivity pattern and collaboration characteristics of authors with multiple retractions, in particular analyzed the pattern and factors explaining for producing multiple retracted publications from the perspective of collaboration.

Meanwhile, although there were surveys and interviews conducted to explore the factors of scientific misconduct among scholars from several countries, such as the need for publications, attitudes of scientific misconduct and institutional policies (Davis, Riske-Morris, & Diaz, 2007; Hofmann, Helgesson, Juth, & Holm, 2015; Mabou Tagne et al., 2020; Mardani, Nakhoda, & Shamsi Gooshki, 2020; Were, Kaguiri, & Kiplagat, 2020; Yi, Nemery, & Dierickx, 2019), these self-report surveys might solely provide a conservative estimation of the emergences of scientific misconduct, considering the concealing to answering sensitive questions among interviewees and other limitations (Fanelli, 2009). Our analysis on the retracted publications will contributed with empirical evidence objectively demonstrating the impact of individual and collaborative factors on scientific misconduct.

The objective of this study is to reveal productivity patterns, collaboration, and reasons explaining why the authors published retracted publications in clinical medicine with empirical evidence, especially the influence of individual scientific careers and collaboration factors on their scientific misconduct. Specifically, the research questions are:

1. Whether the productivity pattern of authors with retracted publications in clinical medicine fit classic Lotka’s law?
2. What are the collaboration characteristics of multi-retracted authors and one-retracted authors?
3. In which stage of scientific careers are researchers more likely to take risks in scientific misconduct?

4. What are the potential motivations and factors which influence the misbehavior of multi-retracted authors?

To further explore the pattern and motivation of producing retracted publications, we selected the top 12 authors in three groups with the most retractions as a case study. The characteristics of group collaboration, and causal factors for misconduct have been discussed. Corresponding policy implications could put forward the governance of scientific misconduct and the improvement of constructions in scientific integrity.

Data and methodology

Searching strategy and data collection

In this study, raw data were retrieved from Science Citation Index Expanded (SCI-EXPANDED) database of Web of Science Core Collection from Clarivate Analytics. We collected retracted publications in clinical medicine from 1978 to 2017 by initially creating a general dataset of retracted publications and then picking out the publications categorized to clinical medicine. The general dataset of retracted publications was obtained using a mixed searching strategy in data retrieval based on the method in our previous study (Zhang, Abraham, & Fu, 2020), namely combing records of retracted publications and retraction notices by the three steps. To generate a sub dataset of clinical medicine from the general dataset, we categorized retracted publications by matching the journal and research fields of each records with Clarivate Analytics Essential Science Indicators (ESI) journal-category list. Clinical medicine with 1309 retracted publications took the largest proportion (27%) of retracted publications among 22 ESI fields.

Authors of retracted publications

All authors referred in retracted publications from the field of clinical medicine were identified according to their corresponding addresses, respectively. Different formats of their names had been unified manually. Except for four anonymous articles, there were 6,161 authors contributing to 1,305 retracted articles in clinical medicine. The "multi-retracted authors" in this article refer to authors who have more than one retracted publication, "one-retracted authors" refer to authors who only published one retracted publication, and the most-retracted authors refer to authors with the most retractions in clinical medicine.

The number of years for one's scientific career in this research count since the year of the author's first scientific publication. To retrieve the first year of the author's publication, we searched the author's name in the official website of the institution for personal curriculum vitae, and databases including PubMed, Google Scholar for the earliest record of the author's publication.

Collaboration indicators of authors

Collaboration was measured as co-authorship between two and more authors. Given different roles of positions of author, first author and corresponding author were discussed. The rate of the first authorship (or corresponding authorship) by the author equals the number of retracted publications first authored (or corresponding authorship) divided by the total number of retracted publications authored by the author.

The analysis of Lotka's law

It is widely acknowledged that most scientific publications are contributed by a few highly productive authors (Holliday, Fuller, Wilson, & Thomas, 2013; Rajgoli &

Laxminarsaiah, 2015). The Lotka's law originally gave a mathematical model measuring scientists' productivity distribution, and has been proved to fit diverse areas including LIS, computer science, medicine, biochemistry, etc. (Holliday et al., 2013; Kawamura, 2009; Pao, 1985; Pinto, Escalona-Fernández, & Pulgarín, 2012; Pulgarín, 2012; Zhang et al., 2018). The application of Lotka's law here is designed to explore the distribution and publication characteristics of multi-retracted authors in comparison to ordinary authors. After using full counting of authors of each paper, this study employed non-linear regression of the power model by SPSS v.22 to calculate the two parameters of Lotka's law. The R-square analysis and K-S statistical test were applied to verify the applicability of Lotka's law in this approach.

Network analysis of the co-authorship

To illustrate the collaboration characteristics of multi-retracted authors and one-retracted authors, an oriented co-authorship network was generated by Gephi 0.9.2. The closeness centrality of the nodes was calculated to quantify the degree of collaboration. The higher closeness centrality an author has, the more influential the author is over the entire network (Yin, Kretschmer, Hanneman, & Liu, 2006).

Results and discussion

Productivity patterns of authors by Lotka's law

Figure 1a reveals the application of Lotka's law of retracted articles in the field of clinical medicine. Figure 1b is the graph on a semi-logarithmic scale to explicit demonstrate the fitness of the model. The non-linear regression model was examined as validly fitted, as the maximum distance between each actual accumulated percentage of authors and expectation was 0.00997, smaller than the K-S test critical value = ${1.63}/{\sqrt{6161}} = 0.02077$. According to the non-linear model, the retracted articles (R^2 = 1.000) matched with Lotka's law model, with exponent n = 3.584, constant C = 0.886.

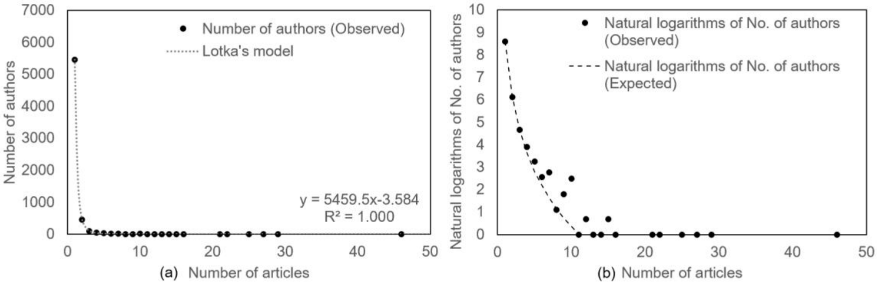


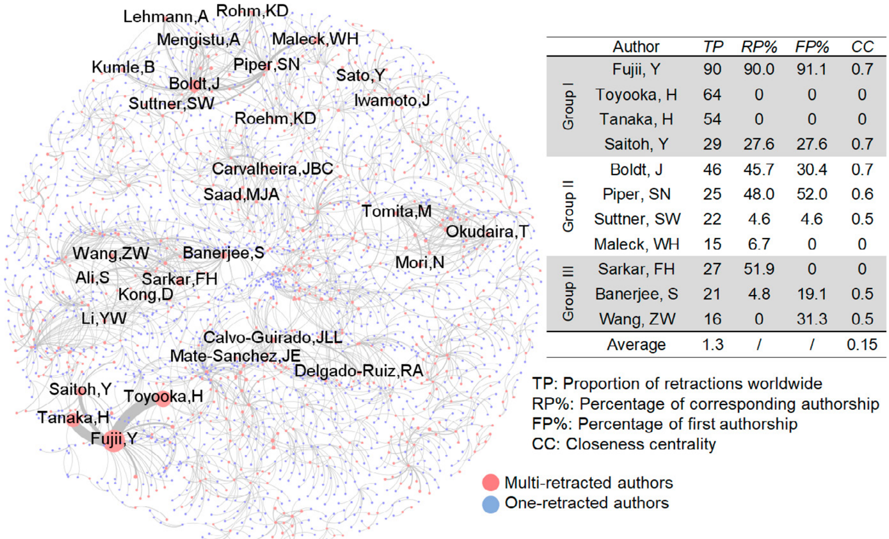
Figure 1. Application of Lotka's law in retractions in the field of clinical medicine

The theoretical Lotka's law has n = 2 and C = 0.6, and parameters varied among different disciplines empirically. In general, as the average level of co-authorship gets lower, such as in social sciences and humanities, parameter n tends to be generally higher; and the larger the C, the lower the author concentration (Kawamura, 2009; Pinto et al., 2012; Rajgoli & Laxminarsaiah, 2015). Previous empirical results in research fields related to clinical medicine varied, such as n = 2.8 and C = 0.52 in critical care medicine (Zhang et al., 2018), n 1.97 ~ 2.52 and C 0.61 ~ 0.77 in radiation oncology, n 2.17 ~ 2.39 and C 0.6622 ~ 0.7205 in dental sciences (Pulgarín, 2012). Notably, both parameters of $n = 3.584$ and $C = 0.886$ of retracted articles in clinical medicine were higher than the theoretical Lotka's law distribution and ordinary medical articles. As high as estimated by $C = 0.886$, most authors, namely 88.6% of all, were retracted only once. Despite a loose collaboration network generally in retracted articles, a small-scaled group of multi-retracted authors was responsible for a disproportionally large number of publications retracted (Holliday et al., 2013).

It is not rare to see such unbalanced or skewed distribution in the scientific community. Considering the potential problems such as errors and scientific misconduct in retracted publications, a predominant aggregation among multi-retracted authors and unbalanced distribution might bring more negative impacts in the scientific community. The following analysis focused on the characteristics of these multi-retracted authors to identify their collaboration patterns and potential motivations for scientific misconduct.

Collaboration at the individual level

Figure 2 revealed the co-authorship network of authors with more than one retracted publication and collaboration characteristics of top 12 authors in three groups. Authors with more than ten retracted articles are presented with name labels in the network. Thick lines indicate the frequency of collaboration. The clockwise bending direction represents the relationship from the first author to the partner.



	Author	*TP*	*RP%*	*FP%*	*CC*
Group I	Fujii, Y	90	90.0	91.1	0.7
	Toyooka, H	64	0	0	0
	Tanaka, H	54	0	0	0
	Saitoh, Y	29	27.6	27.6	0.7
Group II	Boldt, J	46	45.7	30.4	0.7
	Piper, SN	25	48.0	52.0	0.6
	Suttner, SW	22	4.6	4.6	0.5
	Maleck, WH	15	6.7	0	0
Group III	Sarkar, FH	27	51.9	0	0
	Banerjee, S	21	4.8	19.1	0.5
	Wang, ZW	16	0	31.3	0.5
	Average	1.3	/	/	0.15

TP: Proportion of retractions worldwide
RP%: Percentage of corresponding authorship
FP%: Percentage of first authorship
CC: Closeness centrality

● Multi-retracted authors
● One-retracted authors

Figure 2. Co-authorship network and collaboration indicators of authors

There were stronger links among multi-retracted authors in the network. It seems that multi-retracted authors closely collaborated with each other and some were acting as leading authors. The authors who published more retracted publications as first author appeared to have higher closeness centrality than the less published, indicating higher connectivity and more influential control over the retraction fields (Uddin, Hossain, & Rasmussen, 2013). Fujii, Boldt, Sarkar, and Wang in Fig. 2 were known for their high productivity of researches, projects, and quantities of publications (Marcus & Oransky, 2015), once described as leading or incredible example of researchers in the field (Cancer and Metastasis Reviews, 2010; O'Connor, 2008; Oransky, 2010), and accordingly a leading role in the collaborative teams.

Given collaborative occasions, dishonest people usually seek a partner who would also lie, usually called a “partner in crime”, while honest people, who normally didn’t lie but intended to take advantage of their partners’ lies (Gross, Leib, Offerman, & Shalvi, 2018). Consequently, the first offenders were more likely to find acquaintances to commit scientific misconduct (Wray & Andersen, 2018), while co-authors might be willing to take risks or indulge the problematic research practice in consideration of expected benefits. Moreover, mutual confidence and a set of shared norms of behavior between the partners can be constructed by stable co-authorship (Bordons, Aparicio, González-Albo, & Díaz-Faes, 2015).

Scientific careers of multi-retracted authors

To further discover the motivation of most retracted authors, we reviewed their scientific performances during the career and the investigation report on their scientific misconduct by groups, and explored potential motivations for multi-retracted authors’ recidivism of producing retracted publications.

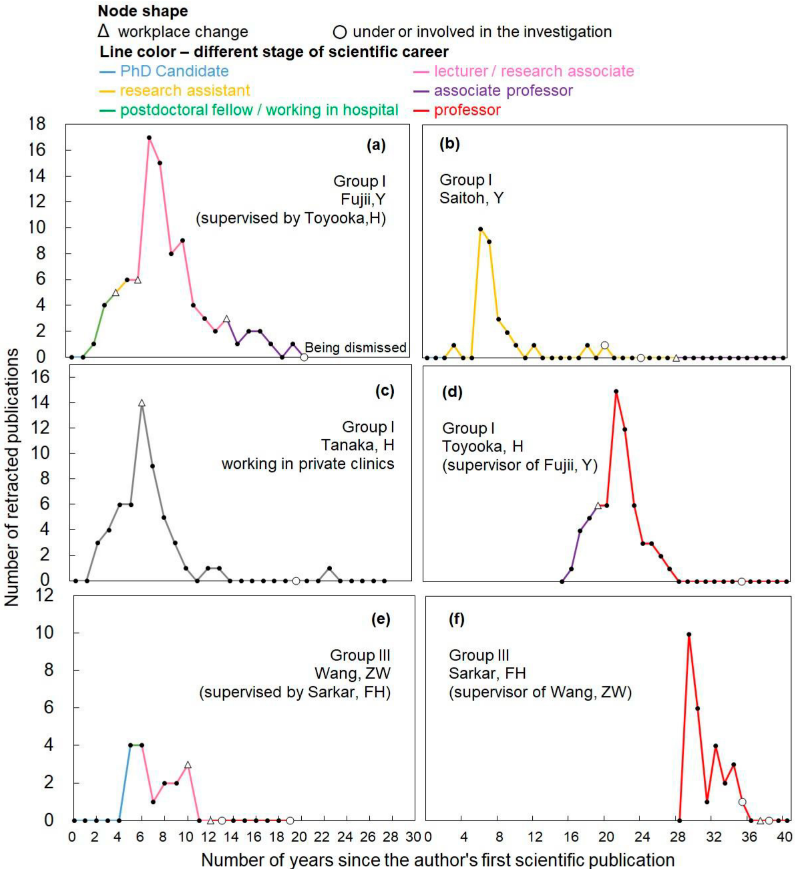


Figure 3. The number of retractions in scientific careers of multi-retracted authors

Figure 3 demonstrated the retraction publication path of retracted articles by top 12 most retracted authors in three groups during their scientific careers. Due to the unavailability of access to some researchers’ resumes on the internet, retractions of six authors during different stages of scientific careers were identified. Distinguishingly two different types of multi-retracted authors were observed: (1) committing scientific misconduct early in their career with a decrease after the peak, and (2) participating in scientific misconduct when mature in their career, namely qualified as senior professor. It is observed that the occurrences of retracted publications are related with individual promotion and mentorship. Firstly, junior researchers appear to have the first retraction quite early in their career, and the rise of retractions is closely associated with getting promoted. For example, Fujii (Fig. 3a) joined the faculty as a research associate from a research assistant, and successfully became a lecturer in the peak year with the most retractions, along with an eight-year decrease of retractions after being named as an associate professor. Wang (Fig. 3e) published four retracted articles in the fifth year as a doctoral undergraduate, which was also the most annual production of retracted articles in his scientific career.

Besides, the rise of the retractions by supervisors also showed consistency with collaborative junior researchers mentored in the same group. Senior researchers could also benefit from publications in pursuit of the promotion as a professor. For instance, Toyooka (Fig. 3d) became the professor in the 19th year near the peak of retractions, similar to Fujii. Moreover, the mutual influence of scientific misconduct though the mentorship seems to be constant in individual career. Apart from the investigation for falsification together with his supervisor Sarkar in the 13th year, Wang (Fig. 3f) was involved in another investigation after becoming a professor elsewhere in the 19th year, according to the latest investigation in 2020.

An in-depth investigation of multi-retracted authors by groups

To give an insight into motivations and factors accounting for scientific misconduct of multi-retracted authors, namely the influences of the mentorship and collaboration, Table 1 summarized the investigation results of most retracted authors based on the related reports (Japanese Society of Anesthesiology, 2012; Justus Liebig University Giessen [JLU], 2011; JLU, 2012; Kuroda, 2012; McCook, 2016; U.S. Office of Research Integrity, 2020). There were three groups under investigation due to scientific misconduct by principal offenders in each group, i.e., Fujii, Boldt, and Sarkar, as well as his postdoctoral fellow, Wang.

Table 1. The in-depth investigation of top 12 multi-retracted authors in three groups

Group	*Author*	*1st publication year*	*1st retraction year*	*Gap*	*Institution*
Group I	Fujii, Y	1991	1993	2	Toho University
	Toyooka, H	1977	1993	16	Teikyo University
	Tanaka, H	1992	1994	2	Toride Kyodo General Hospital
	Saitoh, Y	1992	1995	3	
Group II	Boldt, J	1984	1986	2	Klinikum Stadt Ludwigshafen; Justus Liebig University Giessen (JLU)
	Piper, SN	1997	1999	2	
	Suttner, SW	1998	1999	1	
	Maleck, W	1994	1999	5	
Group III	Sarkar, FH	1977	2006	29	Wayne State University (WSU)
	Banerjee, S	2005	2006	1	
	Wang, ZW	2001	2006	5	
Who calls for investigation & authors' involvement in the investigation? **Group I** Concerns raised by letters from a reader (Christian Apuferu) and request from Joint Editors-in-Chief. Investigated by Japanese Society of Anesthesiology and Toho University. All the authors above were interviewed in 2011. Saitoh was interviewed specifically in 2017.					

Group II Requested by Joint Editors-in-Chief and the State Medical Association of Rheinland-Pfalz, Germany (LÄK-RLP). Investigated by Klinikum Ludwigshafen (2010), LÄK-RLP and JLU (2011). Boldt and Suttner was investigated in 2011. Lack of information about others.

Group III Concerns raised by anonymous letters and from Pubpeer. Investigated by WSU and U.S. Office of Research Integrity (ORI). All were interviewed in 2014. Additionally, Wang was investigated individually in 2020.

Mentorship

Group I Toyooka, H: supervisor of Fujii, Y.

Group II (lack of information)

Group III Sarkar, FH: the head of the lab, Ph. D advisor and postdoctoral supervisor of Wang, ZW.

Results of the investigation & Motivation for the misconduct

Group I

Fujii, Y: falsification of data; lack of ethical approval; authorship falsification. In the pursuit of performance evaluation, obtaining university faculty positions, candidate selection, obtaining public research funding, and applying for social awards.

Toyooka, H: neglected the duty as a supervisor.

Tanaka, H: didn't participate in the falsification and wasn't informed of these publications.

Saitoh, Y: didn't participate in the falsification but agreed to be listed as a co-author.

co-authors above aware of the existence of the publication but left it was in consideration of their scientific output.

Group II:

Boldt, J: lack of ethical approval and patient consent; falsification in publications.

Group III

Sarkar, FH: falsification in grant applications.

Wang, ZW: falsification in the Ph.D. dissertation, publications, and grant applications. Mainly in pursuit of federal grant applications & scientific output. (Although Sarkar himself didn't admit he was motivated.)

Authors' response & punishment

Group I

Fujii, Y: admitted the allegation of scientific misconduct and was finally dismissed by Toho University.

Group II

Boldt, J: was stripped of his title of professor at the JLU for failing to teach. He had been dismissed as a chief physician from Ludwigshafen Hospital in 2010 due to scientific misconduct before. His co-author, Piper once rejected the retraction against the editor with a legal threat.

Group III

Wang, ZW: faced with the withdrawal of Ph.D. degree and a 10-year ban on any federally funded research.

Sarkar, FH: disagreed with the allegation of falsification firstly and sued PubPeer commenters. His employment was rejected by the University of Mississippi due to the event but then he was re-appointed by WSU.

Resulting retractions (in the case report)

Group I *Anesthesia & Analgesia:* 22 are retracted, 24 are suspected of fraud; *Canadian Journal of Anesthesia:* 9 are retracted because of lack of ethical approval.

Group II 88 articles that LAK-RLP was unable to find evidence of IRB approval (Miller, 2011); three including falsification, and were retracted by journals.

Group III 42 publications authored by Sarkar are identified as involved in misconduct and should be retracted; 14 publications authored by Wang should be retracted.

Others

Group I Saitoh was found to have ethical issues in research after being specially investigated by the Japanese Society of Anesthesiologists and quitted the society in 2017.

Group II JLU withdrew the doctoral degree in two cases of former Boldt's doctoral student P. and the doctor T. on account of their misconduct in previous publications and Ph.D. dissertations.

Generally, a common motivation for multi-retracted authors to take risks in scientific misconduct is obtaining research funds and seeking for promotion, which is also reflected in the trend of retracted publications in their scientific careers in Fig. 3. Interestingly, apart from the first authors who should take the main responsibility for misconduct in the investigation, some coauthors would take advantages of the problematic publications considering potential benefits in scientific output. For example, despite the fact that Saitoh didn't participate in the falsification, he also agreed to be listed as a co-author in the retracted publications with Fujii simply in pursuit of scientific output.

Moreover, negative team culture where the duty as supervisor and well-training in research practice is neglected, could facilitate inappropriate research practice and misconduct in the labs. It should be noted that both senior and junior researchers could be influenced by the team culture, especially through mentorship. In both the cases of Group I and II, Toyooka and Sarkar were blamed for neglecting the duty as senior supervisor or PI in the lab. Namely, being reckless and failed to review the data as the supervisor substantially lead to inappropriate preservation or manipulation of data, and miscommunication between laboratory members. The cases here were just a tip of the iceberg, as it was reported that 3/4 of PI in universities don't directly review data (Wright, Titus, & Cornelison, 2008).

Meanwhile, loose regulations and publication policies made it easier for multi-retracted authors to break the rule of scientific norms, namely forged authorship and the lack of ethical approval. Since not all journals require signatures from all authors, it is convenient for authors to fabricate the authorship without others' consent. Fujii, who forged authorship in tens of retracted publications, once confessed that he even didn't need to ask the authorship permission from all the co-authors when submitting the manuscript. Furthermore, irresponsive actions taken by scientific community in respond to the concerns of problematic research would lead to postpone in examination, corresponding investigation and disposal of scientific misconduct. Specifically, journals would fail to respond to the concerns about the problematic research due to the lack of supportive evidence or fear of legal threat from authors (Palus, 2015). On the other hand, corresponding investigation will also be more complex or even intervened under the circumstances where the original data were usually damaged or lost and resistance from the authors. In the case of Sarkar, it took more than 26 months for the investigation committee to find original evidence, under the circumstances where Sarkar and his team have already confused the real data and fabricated data in the lab (McCook, 2016).

Therefore, our results suggest a complete and proactive mechanism that coordinates institutions, journals, and other agencies, to improve transparency and responsibility between collaboration and mentorship, practice on the scientific norms and ethics, responsive actions toward scientific misconduct, and to protect the integrity of future science. More specific and effective examination on publication or research ethics in journal policies are necessary. According to an Editor-in-Chief from *European Journal of Anaesthesiology*, who once commented about the lessons learned from the case of Boldt, after adopting a new rule that requires all articles to provide detailed information about ethics approval, the practice turns out to help editors identify those authors who failed to provide details and never came back to the journal (Tramer, 2011).

On the other hand, corresponding investigation, report, and punishment are vital for building an efficient mechanism to restrain scientific misconduct. As is shown in Fig. 3, there appeared to be a decrease and even termination of publishing retracted articles after being investigated and punished. According to WSU, Sarkar was found to have changed his practice in his lab including reviewing raw data and checking figures ever since the investigation. Since formal sanction and losing jobs are proved to be severe negative results of scientific misconduct perceived by many scholars (Cyranoski, 2014; Pratt, Reisig, Holtfreter, & Golladay, 2019), researchers may reconsider and restrain individual scientific misconduct where the scientific community will respond to misconduct in time.

Acknowledgment

This study is supported by the Soft Science Research Program of the Zhejiang Provincial Department of Science of Technology (No. 2021C35040), and National Natural Science Foundation of China (No. 71804163).

References

Bordons, M., Aparicio, J., González-Albo, B., & Díaz-Faes, A. A. (2015). The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics, 9*(1), 135-144.

Budd, J. M., Sievert, M., Schultz, T. R., & Scoville, C. (1999). Effects of article retraction on citation and practice in medicine. *Bulletin of the Medical Library Association, 87*(4), 437-443.

Cassao, B. D., Herbella, F. A. M., Schlottmann, F., & Patti, M. G. (2018). Retracted articles in surgery journals. What are surgeons doing wrong? *Surgery, 163*(6), 1201-1206.

Cancer and Metastasis Reviews (2010). Biography—Fazlul H. Sarkar, Ph.D. *Cancer and Metastasis Reviews, 29*(3), 379-379.

COPE (2012). *Cooperation between research institutions and journals on research integrity cases: guidance from the Committee on Publication Ethics (COPE).* Retrieved January 30, 2021 from https://publicationethics.org/resources/guidelines-new/cooperation-between-research-institutions-and-journals-research-integrity.

Cyranoski, D. (2014). *Stem-cell pioneer blamed media 'bashing' in suicide note.* Retrieved January 30, 2021 from https://doi.org/10.1038/nature.2014.15715.

Davis, M. S., Riske-Morris, M., & Diaz, S. R. (2007). Causal factors implicated in research misconduct: evidence from ORI case files. *Science and Engineering Ethics, 13*(4), 395-414.

Elia, N., Wager, E., & Tramer, M. R. (2014). Fate of articles that warranted retraction due to ethical concerns: a descriptive cross-sectional study. *PLoS One, 9*(1), e85846.

Fanelli, D. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLoS One, 4*(5), e5738.

Foo, J. Y. (2011). A retrospective analysis of the trend of retracted publications in the field of biomedical and life sciences. *Science and Engineering Ethics, 17*(3), 459-468.

Gross, J., Leib, M., Offerman, T., & Shalvi, S. (2018). Ethical Free Riding: When Honest People Find Dishonest Partners. *Psychological Science, 29*(12), 1956-1968.

Gupta, V., Sahani, A., Mohan, B., & Wander, G. (2013). Negative pressure aerosol containment box: An innovation to reduce COVID-19 infection risk in healthcare workers. *Journal of Anaesthesiology Clinical Pharmacology, 4*(2), 144-147.

Hamilton, D. G. (2019). Continued Citation of Retracted Radiation Oncology Literature-Do We Have a Problem? *International Journal of Radiation Oncology, 103*(5), 1036-1042.

Hofmann, B., Helgesson, G., Juth, N., & Holm, S. (2015). Scientific Dishonesty: A Survey of Doctoral Students at the Major Medical Faculties in Sweden and Norway. *Journal of Empirical Research on Human Research Ethics, 10*(4), 380-388.

Holliday, E., Fuller, C. D., Wilson, L. D., & Thomas, C. R., Jr. (2013). Success breeds success: authorship distribution in the Red Journal, 1975-2011. *International Journal of Radiation Oncology, Biology, Physics, 85*(1), 23-28.

Japanese Society of Anesthesiologists (2012). *Fujii Yoshitaka-shi ronbun ni kansuru chōsa tokubetsu iinkai hōkoku [Report of the Special Committee for Investigation on Yoshitaka Fujii's Treatise]*. Retrieved January 30, 2021 from http://anesth.or.jp/files/download/news/20120629_2.pdf.

Justus-Liebig-Universität [JLU] (2011). *Verstoß gegen gute wissenschaftliche Praxis [Violation of good scientific practice]*. Retrieved January 30, 2021 from https://www.uni-giessen.de/ueber-uns/pressestelle/pm/pm89-11.

Justus-Liebig-Universität [JLU] (2012). *JLU fällt weitere Entscheidung im Boldt-Komplex* [*JLU makes another decision in the Boldt complex].* Retrived January 30, 2021 from https://www.uni-giessen.de/ueber-uns/pressestelle/pm/pm131-12.

Kawamura, C. D. L. T. Y. K. M. (2009). Lotka's law and the pattern of scientific productivity in the dental science literature. *Medical Informatics and the Internet in Medicine, 24*(4), 309-315.

Kuroda, M. (2012). *Disciplinary Decision concerning Dr. Yoshitaka Fujii.* Retrived January 30, 2021 from https://www.toho-u.ac.jp/english/information/march_6_2012.html.

Kuroki, T., & Ukawa, A. (2018). Repeating probability of authors with retracted scientific publications. *Accountability in Research, 25*(4), 212-219.

Lei, L., & Zhang, Y. (2018). Lack of Improvement in Scientific Integrity: An Analysis of WoS Retractions by Chinese Researchers (1997-2016). *Science and Engineering Ethics, 24*(5), 1409-1420.

Luther, F. (2010). Scientific misconduct: Tip of an iceberg or the elephant in the room? *Journal of Dental Research, 89*(12): 1364–1367.

Mabou Tagne, A., Cassina, N., Furgiuele, A., Storelli, E., Cosentino, M., & Marino, F. (2020). Perceptions and Attitudes about Research Integrity and Misconduct: A Survey among Young Biomedical Researchers in Italy. *Journal of Academic Ethics, 18*(2), 193-205.

Marcus, A., & Oransky, I. (2015). *How the Biggest Fabricator in Science Got Caught*. Retrieved January 31, 2021 from http://nautil.us/issue/24/error/how-the-biggest-fabricator-in-science-got-caught.

Mardani, A., Nakhoda, M., & Shamsi Gooshki, E. (2020). Relationship among factors affecting research misconduct in medical sciences in Iran. *Accountability in Research*, *27*(7), 1-27.

McCook, A. (2016). *Details of investigative report into Sarkar released by ACLU.* Retrieved December 1, 2020 from: https://retractionwatch.com/2016/11/17/details-of-investigative-report-into-sarkar-released-by-aclu/.

McHugh, U. M., & Yentis, S. M. (2019). An analysis of retractions of papers authored by Scott Reuben, Joachim Boldt and Yoshitaka Fujii. *Anaesthesia, 74*(1), 17-21.

Miller, D. (2011). Retraction Note to: Diltiazem may preserve renal tubular integrity after cardiac surgery. *Canadian Journal of Anaesthesia, 58*(9), 881, 881-882.

Mistry, V., Grey, A., & Bolland, M. J. (2019). Publication rates after the first retraction for biomedical researchers with multiple retracted publications. *Accountability in Research, 26*(5), 277-287.

Mongeon, P., & Larivière, V. (2016). Costly collaborations: The impact of scientific fraud on co-authors' careers. *Journal of the Association for Information Science and Technology, 67*(3), 535-542.

O'Connor, J. (2008). *Wayne State researcher receives award for potential prostate cancer treatment.* Retrieved January 29, 2021 from: https://today.wayne.edu/news/2008/05/29/wayne-state-researcher-receives-award-for-potential-prostate-cancer-treatment-2908.

Oransky, I. (2010). *After misrepresentation allegations, German anesthesiologist Joachim Boldt out as hospital's chief physician.* Retrieved January 30, 2021 from https://retractionwatch.com/2010/11/26/after-misrepresentation-allegations-german-anesthesiologist-joachim-boldt-out-as-hospitals-chief-physician.

Palus, S. (2015). *Boldt's retraction count upped to 94, co-author takes legal action to prevent 95th*. Retrieved December 1, 2020 from https://retractionwatch.com/2015/10/12/boldts-retraction-count-upped-to-94-co-author-takes-legal-action-to-prevent-95th.

Pao, M. L. (1985). Lotka's law: A testing procedure. *Information Processing & Management, 21*(4), 305-320.

Pinto, M., Escalona-Fernández, M. I., & Pulgarín, A. (2012). Information literacy in social sciences and health sciences: a bibliometric study (1974–2011). *Scientometrics, 95*(3), 1071-1094.

Pratt, T. C., Reisig, M. D., Holtfreter, K., & Golladay, K. A. (2019). Scholars' preferred solutions for research misconduct: results from a survey of faculty members at America's top 100 research universities. *Ethics & Behavior, 29*(7), 510-530.

Pulgarín, A. (2012). Dependence of Lotka's law parameters on the scientific area. *Malaysian Journal of Library & Information Science, 17*(1), 41-50.

Rajgoli, I. U., & Laxminarsaiah, A. (2015). Authorship pattern and collaborative research in the field of spacecraft technology. *The Electronic Library, 33*(4), 625-642.

Sabir, H., Kumbhare, S., Parate, A., Kumar, R., & Das, S. (2015). Scientific misconduct: a perspective from India. *Medicine, Health Care and Philosophy, 18*(2), 177-184.

Saikia, P., & Thakuria, B. (2019). Retraction of papers authored by Yuhji Saitoh - Beyond the Fujii phenomenon. *Indian journal of anaesthesia, 63*(7), 571-584.

Stavale, R., Ferreira, G. I., Galvao, J. A. M., Zicker, F., Novaes, M., Oliveira, C. M., & Guilhem, D. (2019). Research misconduct in health and life sciences research: A systematic review of retracted literature from Brazilian institutions. *PLoS One, 14*(4), e0214272.

Steen, R. G. (2011). Retractions in the medical literature: how many patients are put at risk by flawed research? *Journal of Medical Ethics, 37*(11), 688-692.

Tang, L., Hu, G., Sui, Y., Yang, Y., & Cao, C. (2020). Retraction: The "Other Face" of Research Collaboration? *Science and Engineering Ethics, 26*(3), 1681-1708.

Tramer, M. R. (2011). The Boldt debacle. *European Journal of Anaesthesiology, 28*(6), 393-395.

Uddin, S., Hossain, L., & Rasmussen, K. (2013). Network effects on scientific collaborations. *PLoS One, 8*(2), e57546.

U.S. Office of Research Integrity [ORI] (2020). *Case Summary: Wang, Zhiwei*. Retrieved January 30, 2021 from https://ori.hhs.gov/content/case-summary-wang-zhiwei.

Were, E., Kaguiri, E., & Kiplagat, J. (2020). Perceptions of occurrence of research misconduct and related factors among Kenyan investigators engaged in HIV research. *Accountability in Research, 27*(6), 372-389.

Wray, K. B., & Andersen, L. E. (2018). Retractions in Science. *Scientometrics, 117*(3), 2009-2019.

Wright, D. E., Titus, S. L., & Cornelison, J. B. (2008). Mentoring and research misconduct: an analysis of research mentoring in closed ORI cases. *Science and Engineering Ethics, 14*(3):323-36.

Yi, N., Nemery, B., & Dierickx, K. (2019). Perceptions of research integrity and the Chinese situation: In-depth interviews with Chinese biomedical researchers in Europe. *Accountability in Research, 26*(7), 405-426.

Yin, L.-Z., Kretschmer, H., Hanneman, R. A., & Liu, Z.-Y. (2006). Connection and stratification in research collaboration: An analysis of the COLLNET network. *Information Processing & Management, 42*(6), 1599-1613.

Zhang, M., & Grieneisen, M. (2012). The impact of misconduct on the published medical and non-medical literature, and the news media. *Scientometrics, 96*(2), 1–15.

Zhang, Q., Abraham, J., & Fu, H.-Z. (2020). Collaboration and its influence on retraction based on retracted publications during 1978-2017. *Scientometrics, 125*(1), 213-232.

Zhang, Z., Van Poucke, S., Goyal, H., Rowley, D. D., Zhong, M., & Liu, N. (2018). The top 2,000 cited articles in critical care medicine: a bibliometric analysis. *Journal of Thoracic Disease, 10*(4), 2437-2447.

Enhanced Author Bibliographic Coupling Analysis Using Syntactic and Semantic Citation Information

Ruhao Zhang[1] and Junpeng Yuan[2]

[1] *zhangruhao@mail.las.ac.cn*

[2] *yuanjp@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing (China)
Department of Library, Information and Archives Management, School of Economics and Management,
University of Chinese Academy of Sciences, Beijing (China)

Abstract

Author Bibliographic Coupling Analysis (ABCA) is the extension of Bibliographic Coupling theory at the author level, widely used in mapping intellectual structure and scholar communities. However, the assumption of equal citations and the complete dependence on explicit counts may affect its effectiveness in today's complex context of the discipline development. This research proposes a new approach that uses multiple full-text data to improve ABCA called Enhanced Author Bibliographic Coupling Analysis (EABCA). By mining the syntactic and semantic information of citations, the new approach considers more diverse and in-depth dimensions as the basis of author bibliographic coupling strength. Comparative empirical research was then conducted in the field of oncology. Results show that the new approach can more accurately reveal the relevance between authors and map a more detailed domain intellectual structure.

Introduction

The Bibliographic Coupling (BC) concept originated from Kessler's (1963) discovery: When two documents are more similar in topic, they share more common references. Zhao & Strotmann (2008a) extended BC and proposed Author Bibliographic Coupling Analysis (ABCA), which is now used to characterize the research interest of communities of current active authors and reveal a field's intellectual structure. It has been proven that ABCA is a supplement to Author Co-citation Analysis (ACA), especially in finding active authors and research fronts (Ma, 2012; Zhao & Strotmann, 2014).

However, ABCA, like most bibliometric methods based on Citation Analysis (CA), has inherent flaws. CA relies on the assumption that all citations are equivalent, which ignores the different deep meanings represented by citations (Bornmann & Daniel, 2008), and only uses simple count as the basis for relevance strength. The count-based strength may affect the quality of utility of these methods, including ABCA, in the complex background where interdisciplinary research is increasing and rapidly progressing.

Fortunately, the accessibility of full-text scientific articles in databases such as PubMed Central (PMC), Citeseer, and arXiv, promotes the emergence of Content-based Citation Analysis (CCA). CCA extends the scope of CA from metadata to information inside scientific articles, and studies the motivations, sentiments, and topics of citation, which further provides strong support for scientific evaluation, retrieval, recommendation, automatic summarization, etc. CCA has been regarded as the next generation of CA (Ding et al., 2014). The current research of CCA can be summarized on the syntactic and semantic level. The syntactic level focuses on external characteristics of citations such as mention frequency in one paper (Count-X) (Ding et al., 2013), position (Gipp & Beel, 2009; Liu & Chen, 2012; Habib & Afzal, 2019). The semantic level focuses on the internal content of citation (Liu, Zhang, & Guo, 2013; Kumar et al., 2017). At present, researchers have engaged in content-based improvement for ACA using co-citation position (An et al., 2017) and co-citation sentence (Jeong & Song, 2014; Kim et al., 2016). These studies have achieved encouraging results. However, the research on ABCA is still almost blank.

Can CCA also take effect in ABCA? This investigation proposes a new model that uses full-text information to improve ABCA, named Enhanced Author Bibliographic Coupling Analysis (EABCA), to explore the feasibility and effectiveness of expanding the depth of data in the commonly used author-based bibliometric methods. Unlike ABCA, EABCA focuses on the similarities of citing behaviors between authors in multiple dimensions (Figure 1), comparing contents, locations, and mention frequency of citations, to obtain comprehensive relevance strength and then assign each coupling a different weight.

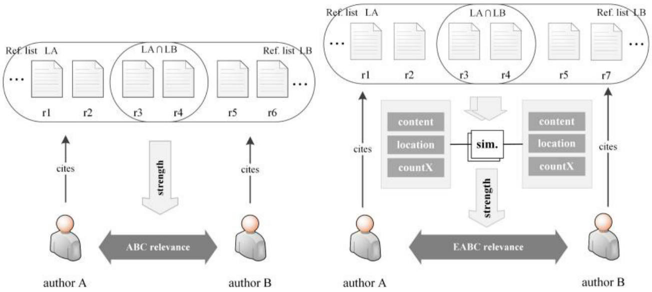
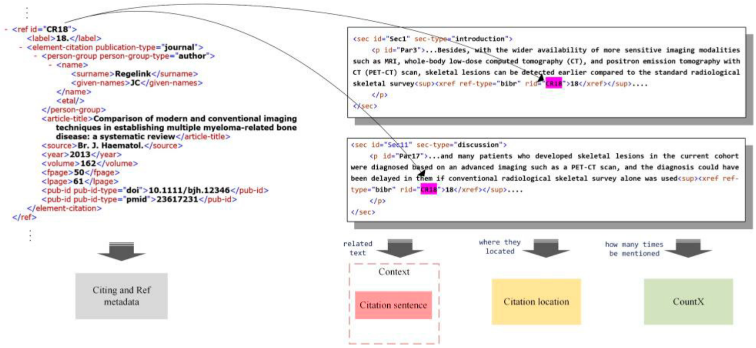


Figure 1. Difference between ABCA (left) and EABCA (right)

EABCA Model Design

In this section, we give an introduction to EABCA's three essential parts.

Full-text Data Extraction



```
      ⋮
- <ref id="CR18">
    <label>18.</label>
  - <element-citation publication-type="journal">
    - <person-group person-group-type="author">
      - <name>
          <surname>Regelink</surname>
          <given-names>JC</given-names>
        </name>
        <etal/>
      </person-group>
      <article-title>Comparison of modern and conventional imaging
        techniques in establishing multiple myeloma-related bone
        disease: a systematic review</article-title>
      <source>Br. J. Haematol.</source>
      <year>2013</year>
      <volume>162</volume>
      <fpage>50</fpage>
      <lpage>61</lpage>
      <pub-id pub-id-type="doi">10.1111/bjh.12346</pub-id>
      <pub-id pub-id-type="pmid">23617231</pub-id>
    </element-citation>
  </ref>
      ⋮
```

```
<sec id="Sec1" sec-type="introduction">
    <p id="Par3">...Besides, with the wider availability of more sensitive imaging modalities
such as MRI, whole-body low-dose computed tomography (CT), and positron emission tomography with
CT (PET-CT) scan, skeletal lesions can be detected earlier compared to the standard radiological
skeletal survey<sup><xref ref-type="bibr" rid="CR18">18</xref></sup>....
    </p>
</sec>
```

```
<sec id="Sec11" sec-type="discussion">
    <p id="Par17">...and many patients who developed skeletal lesions in the current cohort
were diagnosed based on an advanced imaging such as a PET-CT scan, and the diagnosis could have
been delayed in them if conventional radiological skeletal survey alone was used<sup><xref ref-
type="bibr" rid="CR18">18</xref></sup>....
    </p>
</sec>
```

Figure 2. Procedure of full-text data extraction

Table 1. Example of the extracted data structure

CitingAuthor	'Lacy,Martha Q.'
CitedArticle	' Comparison of modern and conventional imaging techniques in… '
Citedpmid	23617231
CitingArticle	'Risk stratification of smoldering multiple myeloid…'
Citingpmc	5997745
Citingpmid	29895887
Citation Content	[' …Besides, with the wider availability of more sensitive imaging modalities such as MRI…', '…and many patients who developed skeletal lesions in…']
Location	[0.050, 0.856]
Count-X	2

Extraction of full-text data is essential preliminary work for our algorithm. Take the XML format data provided by PMC as an example (Figure 2). For each paper, the tag '<xref>' is used to locate the citations for each reference. Subsequently, citations' data of content, location, and Count-X, as well as other metadata, will be parsed. Finally, for each paper' reference, the data structure shown in Table 1 will be formed for each collaborator. Note that the citation contents include the citation sentences and their adjacent sentences separated from other ones by full stops (not dot etc.), and the locations are obtained by dividing the sequential index of each citation's first word by the length of the paper.

Language Model Pre-trainning

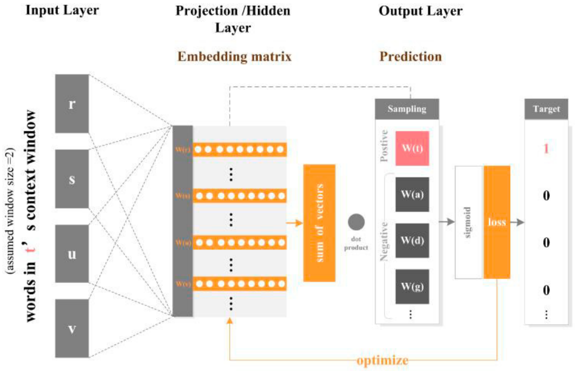


Figure 3. The mechanism of Continuous Bag of Words (CBOW)

Since EABCA involves calculating the content similarity between citations, Natural Language Processing techniques are required to provide word representation support. To overcome the disconnected relationship between word form and semantic using TF-IDF as previous studies (Kim et al., 2016), considering the efficiency and scalability, word2vec was chosen for pre-training distributional representation of words based on the field corpus. The word2vec algorithm uses a neural network to learn semantic and contextual associations of words from a large corpus in an unsupervised way (Mikolov et al., 2013). Continuous Bag of Words (CBOW) is one of its two modes, which has weaker sensitivity to rare words and is more suitable for our large-scale scientific literature corpus modeling than the other. As shown in Figure 3, CBOW predicts the central word t by inputting the sum of embedding vectors of words from t's context window and then backpropagates the loss to optimize the parameters in the embedding matrix. After training multiple iterations on the entire corpus dataset, those words with similar contexts tend to have similar embedding vectors.

Table 2. Example of the most similar words of 'glioma' in the trained language model

word	*similarity*	*word*	*similarity*
GBM	0.7782	astrocytomas	0.5494
glioblastoma	0.7727	GSCS	0.5428
astrocytoma	0.6234	GC	0.5425
osteosarcoma	0.5787	PDAC	0.5407
HCC	0.5721	CCRCC	0.5225

To fit the word2vec model to the terminology features and relations of the specific domain in advance, we pre-processed all full-text articles in our dataset and then utilized them as a corpus in pre-training. Table 2 shows an example of the most similar ten words of 'glioma' in the trained model. The close correlations between glioma and these words, such as GBM (glioblastoma), are correctly represented. In other words, even if two sentences on the same topic use completely different term forms, we can still compute the sentences' similarity.

Calculation

Consider two authors A and B. P_A and P_B denote the sets of articles of A and B, while L_A and L_B denote the reference sets of A and B. Suppose an intersection $L_{AB} = L_A \cap L_B$ exists, where we have eliminated those references included in A and B's co-author articles. For each reference $r \in L_{AB}$, We aggregate all the citation contents and locations of r in P_A and P_B, generating the total contents list donated by CON_A and CON_B, and the total location list donated by LOC_A and LOC_B. We then sum all the Count-X of r in P_A and P_B, generating the total Count-X donated by COX_A and COX_B.

Based on the above, the similarity between A and B when citing r is defined as (1).

$$Similarity_r(A,B) = cos\begin{pmatrix} w_1 \cdot \arccos(CON\_sim) + \\ w_2 \cdot \arccos\,(LOC\_sim) + \\ w_3 \cdot \arccos\,(COX\_sim) \end{pmatrix} \quad (1)$$

$$(w_1 + w_2 + w_3 = 1, w_1 \in [0,1], w_2 \in [0,1], w_3 \in [0,1])$$

A linear combination of the similarities' angles is adopted in (1) considering different nature of combined data suggested by Glänzel & Thijs (2011). The parameters, w_1, w_2 and w_3, are given to CON_sim, LOC_sim, and COX_sim, which denote the similarities of CON, LOC, and COX, respectively. After every $r \in L_{AB}$ has been traversed and calculated by (1), we added all the $Similarity_{rn}(A,B)$ to obtain the EABC relevance strength between author A and B as

$$EABC\_Relevance(A,B) = \textstyle\sum_1^n Similarity_{rn}(A,B) \quad (2)$$

The three sub-formulas in (1) are defined as follows:

- CON_sim

The items in $CON_A[con_{A1}, con_{A2}, \dots, con_{Am}]$ and $CON_B[con_{B1}, con_{B2}, \dots, con_{Bn}]$ are enumerated paired. For each pair $p(con_{Am}, con_{Bn})$, i dimension term vectors are extracted from the trained word2vec model and summed to generate sentence vectors V_A and V_B . We then apply the cosine to measure their similarity (3):

$$sim_{mn}(con_{Am}, con_{Bn}) = \cos\,(V_A,\ V_B) = \frac{\sum_i V_{Ai} V_{Bi}}{\sqrt{\sum_i V_{Ai}^2}\sqrt{\sum_i V_{Bi}^2}} \quad (3)$$

After obtaining all $sim_{mn}(p)$, we average their value as the content semantic similarity denoted by CON_sim. Note that the CON_sim value is also used to support identifying content-irrelevant relations. If A and B's CON_sim ranks in the lowest 20% of all CON_sim values in the dataset, their relation will be ignored.

- LOC_sim

The items in $LOC_A[loc_{A1}, loc_{A2}, \dots, loc_{Am}]$ and $LOC_B[loc_{B1}, loc_{B2}, \dots, loc_{Bn}]$are enumerated paired. We refer to the generalized alternative of the Jaccard Index (Ružička, 1958), for each pair $p(loc_{Am}, loc_{Bn})$, the smaller value is divided by the larger one as $\min(p)$ / $\max\,(p)$ into a similarity ratio sim_{mn} represented how similar this pair of locations are. The set $sim = \{sim_1, sim_2, \dots, sim_{mn}\}$ is then averaged as the location similarity LOC_sim.

- COX_sim

COX_A and COX_B are first normalized by the number of A's and B's total references, respectively. We then use the larger value to divide the smaller one to obtain COX_sim, which represents how similar the importance of r is for author A and B. Note that if both COX_A and COX_B are only 1, the calculation will not be performed because in this case COX_sim will incorrectly be high, and theoretically, we cannot further deduce the two authors are similar from the fact that r is unimportant to them.

Research Design and Methods

Returning to the research question, we conduct the empirical research as shown in Figure 4 to verify the effect of EABCA. We first collect and process the full-text data from PMC in a specific field, then carry out a comparative analysis on the two models, and present the results through factor analysis and network analysis.

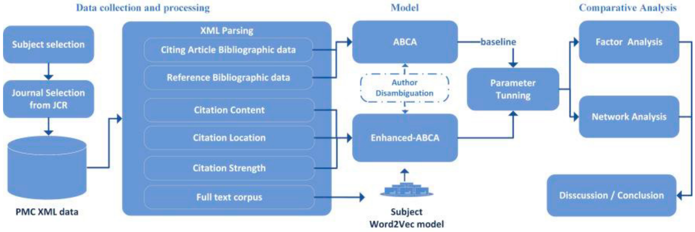


Figure 4. Framework of the research process

Data collection

The field of Oncology was chosen for our empirical study, in which the top 15% (36 items) journals ranking by Impact Factor in the Journal Citation Reports 2019 were selected, and the full-text XML data of these journals during 2016-2020 was downloaded from PMC. Considering the unavailability of open access in some journals and the difference in citation pattern of Review with Research articles, 7,056 full-text research articles from 33 journals were finally retained. Table 3 shows the details of a part that provides more than 400 full-text research articles of these selected source journals.

Table 3. List of source journals providing full-text research paper (part)

Journal title	*IF*	*Full-text research papers*	*% of all published*
Journal of Experimental & Clinical Cancer Research	32.956	1235	84.01%
Oncogene	7.971	990	90.49%
Journal for Immunotherapy of Cancer	11.577	772	63.96%
Molecular Oncology	6.574	541	85.06%
Oncoimmunology	5.869	538	81.64%
Leukemia	13.357	470	70.15%
Molecular Cancer	15.302	411	51.89%

After processing these data in a full-authors pattern, we have 36,532 authors, who totally have 1,741,626 reference entries. Using the full-author pattern in such a field with dense collaboration, which has an average of more than 12 authors and a maximum value of up to 507, generates large-scale redundant data. Therefore, to focus more on the active authors and significant intellectual structure during 2016-2020, and apply quality control well i.e. author disambiguation, thresholds on the number of publications were adopted.

Methodology

Author Bibliographic Coupling Analysis (ABCA)

ABCA is used as the theoretical basis and the comparison in this study. Currently, there are three methods for its strength calculation (Ma, 2012), including the simple method (Rousseau, 2010), minimum method (Zhao & Strotmann, 2008a), and combined method (Leydesdorff, 2011). We chose the minimum method as the comparison: for each reference r in author A and B's reference intersection set, it records the frequency w of A and B respectively citing r in the weighted reference sets W_A and W_B, and sum all the minimum w to obtain the strength.

Author Name Disambiguation (AND)

Author name ambiguity is a crucial problem in author-based bibliometric studies. We adopted a procedure of balancing efficiency and quality to solve the prominent cases as follows:

- Multiple author entities share a single name. We checked each name in the filtered list to confirm whether its owner belongs to multiple affiliations and whether its owner's affiliation repeatedly changes within a period of 2 years. If so, and no entity owned more than 90% of this name's articles, we removed the name (i.e., Wang, Wei).

• Single author entity with multiple name forms. Based on the above, we performed similar name pattern matching for authors and then verify whether these names belong to the same entity by affiliations and research topics. If so, we respectively merge their reference sets and publication sets. Some author entities with multiple forms identified are shown in Table 4.

Table 4. Example of author entities with different name forms

Author entity	*Different name forms*
Munshi,Nikhil C.	['Munshi,Nikhil C.', 'Munshi,Nikhil', 'Munshi,N C', 'Munshi,Nikhil C', 'Munshi,NC']
Anderson,Kenneth C.	['Anderson,Kenneth C.', 'Anderson,K C', 'Anderson,Kenneth C', 'Anderson,Kenneth']
Kaufman,Howard L.	['Kaufman,Howard L.', 'Kaufman,Howard L', 'Kaufman,H. L.', 'Kaufman,H.']

Parameter Tuning

To fit the EABCA model to the characteristics of citation in Oncology, we conducted a parameter tuning toward w_1,w_2 & w_3 in (1) on 66 different weight configurations. The tuning was based on 218 top authors who published more than five papers for better control.

To evaluate the clustering solutions, we applied the method using Medical Subject Headings (MeSH) introduced by Zhang, Xu & Zhang (2019) toward the author level. MeSH is a controlled vocabulary of terms annotated by specialists in life sciences. For each author, we associate his papers with MeSH descriptors in PubMed. After eliminating general and low-frequency terms, we aggregated descriptors and their frequency weights to form the author's MeSH Pool. Subsequently, we used the Jaccard Index, defined as

$$J(A,B) = \frac{|A \cap B|}{|A|+|B|-|A \cap B|} = \frac{|\sum \min(w_{cA}, w_{cB})|}{|Pool_A|+|Pool_B|-|\sum \min(w_{cA}, w_{cB})|} \quad (4)$$

where $Pool_A$ and $Pool_B$ denote author A's and B's MeSH Pools, w_{cA} and w_{cB} denote the weights of descriptor c ∈ $Pool_A \cap Pool_B$ in frequency-weighted sets of A and B, respectively. If $J(A,B)$ exceeds a certain threshold, A and B were considered to belong to the same category. After experimentation and expert judgment, the threshold was set to 0.1.

The matrix generated by EABCA and ABCA was clustered using K-means after converting into the Pearson correlation matrix. Based on (4) and the process shown in Table 5, the Precision, Recall and F1-score is calculated as (5), (6) and (7). We set the number of clusters for K-means to 10 because both EABCA ($w_1 = 0.33$,$w_2 = 0.33$,$w_3 = 0.33$) and ABCA have relatively good performance.

Table 5. Process to evaluate the result of each pair of authors

Two authors belong to the same cluster?	***Two authors belong to the same category based on their MeSH Pool?***	
	Yes	*No*
Yes	True Positive (TP)	False Positive (FP)
No	False Negative (FN)	True Negative (TN)

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1 - score = \frac{2Precision \cdot Recall}{Precision+Recall} \quad (7)$$

Factor Analysis

Factor analysis is a dimension reduction technique commonly applied in author-based bibliographic analysis for exploring the implied relevance between domain representative authors (White & McCain, 1998). In this study, the analysis was implemented following the steps: (a) Selecting and disambiguating 218 top authors who published more than five papers. (b) Building matrixes with mean as the diagonal, including one based on the BC frequency for

ABCA and one on the EABC relevance for EABCA, which were both then converted to correlation matrixes. (c) Extracting factors based on observations of eigenvalues, scree plot, and percentage of variance. (Hair et al., 2009) (d) Applying oblimin rotation for better interpreting. (e) Labelling the factors and comparing the results.

Network Analysis

To study on a larger scale and intuitively map the relevance between authors and research interest groups, we conducted a further network analysis using Gephi on 723 authors who published more than three papers. The edge weight depended on BC frequency and EABC relevance for ABCA and EABCA, respectively. Louvain algorithm (Blondel et al., 2008) was utilized for community discovery, which is highly efficient facing large networks.

Result

Parameter Tuning Result

Figure 5 shows the performances of models with different parameter values. By comparing the values, we notice that the parameter w_1 has a strong positive correlation with Precision (0.798 with significance at 0.01), and w_2 has a weak positive correlation with Recall (0.251). On the contrary, there exists a negative correlation between w3 and Precision (-0.709) and Recall (-0.647). The above indicates that the content and location similarity switched by w_1 and w_2 play roles in improving the performance, while a too high proportion of Count-X similarity may not lead to an outstanding result because the w_1 and w_2 are therefore reduced.

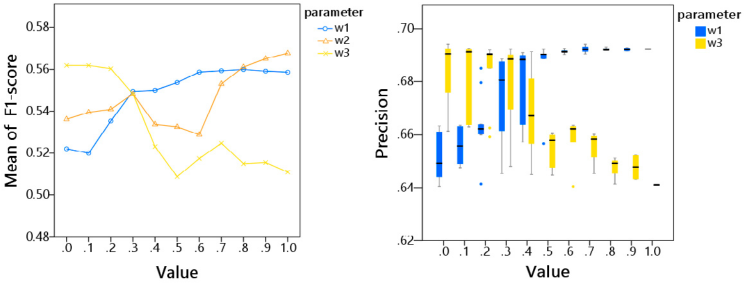


Figure 5. F1-score Mean Plot of three parameters (left) and Precision Box-plot of parameter w_1 and w_3(right)

Table 6 shows the performance of the top 5 models ranked by F1-score, the model with the highest Precision and Recall, and ABCA in parameter tuning. These scores are all average values generated by 30 rounds of reproducible initialized centroids to ensure the stability of the results. The Recall of the top model significantly exceeds ABCA by 10.961%, while the Precision increases by 6.148% and the F1-score by 9.887%. To a certain extent, the above indicates the advantage of EABCA over ABCA from a quantitative perspective. Hence, the EABCA ($w_1 = 0.4$, $w_2 = 0.4$, $w_3 = 0.2$) was chosen.

Table 6. Performance of models on K-means clustering

Rank	*Model*	*Precision*	*Recall*	*F1-score*
1	EABCA(w1=0.4,w2=0.4,w3=0.2)	69.033%	48.971%	**57.268%**
2	EABCA(w1=0.3,w2=0.4,w3=0.3)	68.788%	48.814%	57.064%
3	EABCA(w1=0.3,w2=0.6,w3=0.1)	68.738%	48.722%	56.993%
4	EABCA(w1=0.3,w2=0.5,w3=0.2)	68.761%	48.539%	56.877%
5	EABCA(w1=0.4,w2=0.5,w3=0.1)	69.089%	48.244%	56.796%
7	EABCA(w1=0.0,w2=1.0,w3=0.0)	66.336%	**49.624%**	56.774%
15	EABCA(w1=0.7,w2=0.3,w3=0.0)	**69.417%**	47.301%	56.250%
66	Baseline - ABCA	62.885%	38.010%	47.381%

Factor Analysis Result

After name disambiguation, 218 authors published more than five papers were retained for factor analysis. Ten factors were extracted, explaining 70.300% of the total variance for ABCA, while 13 factors were extracted, explaining 70.389% of the total variance for EABCA. Details are shown in Table 7. Note that an author variable may have loadings on many factors here; therefore, our factor analysis aims at comparing the differences between the two models in mining latent topics rather than giving each author a one-to-one category.

The topic labels in Table 7 come from our keyword extraction procedure for each factor. For the coupling authors of every factor, the titles and MeSH descriptors were extracted from their BC articles to support topic identification based on TF-IDF. We then surveyed the representative authors online to label factors with the assistance of domain experts.

Table 7. Factor Analysis results of ABCA and EABCA

Topic Label	*ABCA (10 Factors,70.300 %)*			*EABCA (13 Factors, 70.389%)*		
	Factor	*Size*	*Highest loading*	*Factor*	*Size*	*Highest loading*
RNA Interference	F1, F7, F8	114	0.951	F1	30	0.933
Multiple Myeloma and Therapy	F4	28	1.008	F2	28	1.016
Immunotherapy	F3	26	0.917	F3	24	0.929
Multiple Myeloma and Genomics	F2	27	0.692	F4	25	1.006
Stem cell Transplantation	F5, F9	35	0.896	F5, F10	36	0.931
Acute Myeloid Leukemia	F6	25	0.753	F6	18	0.839
Gene Expression Regulation in Esophagogastric & Stomach Neoplasms				F7	24	0.729
Gastric & Colorectal Neoplasms and Carcinogenesis Target Research				F8	15	0.774
Myeloproliferative Disorders	F10	20	0.825	F9	14	0.904
Lymphoma				F11	11	0.680
Signaling Pathways / Computational Biology / Head and Neck Neoplasms				F12	9	0.539
Gene Expression Regulation in Liver Neoplasms				F13	33	0.761

Table 8. Factors exclusively identified by EABCA

Factor	*keywords from MeSH and Title(by Tf-idf)*	*Representative author (with loading in structure matrix)*
F7	Stomach Neoplasms / Gene Expression Regulation, Neoplastic / Esophagogastric Junction / Signal Transduction / MicroRNAs	Shen, Lin (0.716) Xu, Zekuan (0.545)
F8	HCT116 Cells / Colorectal Neoplasms / Caco-2 Cells / HT29 Cells / E1A-Associated p300 Protein / rela / last2yap1/ p300-mediated	To,Ka Fai (0.828) Wong,Chi Chun (0.757)
F11	Lymphoma, Large B-Cell, Diffuse / DNA Copy Number Variations / Lymphoma / Rituximab / Doxorubicin / dlbcl	Nowakowski,Grzegorz S. (0.654) Witzig,Thomas E. (0.628)
F12	ErbB Receptors / Computational Biology / Gene Expression Regulation, Neoplastic / Head and Neck Neoplasms / Epidermal Growth Factor / histone / tmem16a / six3	Teng, Yong (0.519) Liu, Changhong (0.516)
F13	Gene Expression Regulation, Neoplastic / Carcinoma, Hepatocellular / Liver Neoplasms / MicroRNAs / Epithelial-Mesenchymal Transition / m6a	Tu, Kangsheng (0.849) Zhou, Lin (0.705)

From the results, we find that ABCA and EABCA are similar in some of the identified factors, including F2, F3, and F4. The main difference occurs in F7, F8, F11, F12, and F13, which are only identified in EABCA. These factors' highest loadings are not high, and some of them closely associate with others, i.e., F7, F8, F13, and F1, while F11 and F12 are relatively independent. However, the author groups under these new factors have their own common characteristics in research content or scenarios, as shown in Table 8. Factor F7 & F13 both involve the studies of RNA interference, but F7 focuses on esophagogastric & stomach neoplasms while F13 on liver-related neoplasms. F8 involves gastric & colorectal neoplasms, and it focuses on carcinogenesis targets. F11 involves lymphoma, including its clinical therapy as well as related basic genomic research. F12 mainly involves researches on signaling pathways, of which most use head and neck neoplasms as specific objects.

Network Analysis Result

After name disambiguation, 723 authors who published more than three papers were retained for network analysis. For these authors, ABCA generated 73,984 edges, EABCA generated 40,878 edges. The overview of networks is shown in Table 9, where we find that EABCA's network is sparser and has weaker connectivity.

Table 9. Overview of network of ABCA and EABCA

	Num. of edges	*Average Degree*	*Average Weighted Degree*	*Density*	*Average Path length*	*Connected Components*
ABCA	73984	204.658	570.866	0.283	1.723	2
EABCA	40878	113.079	144.191	0.157	1.944	3

To show a distinctive network structure, we adopted edge filtering for both ABCA and EABCA. ABCA retained 21,677 edges with a weight of 3 or more, while EABCA retained 9,304 edges with a weight of 1.5 or more because of its relatively lower edge weight. The community discovery algorithm found five modules for ABCA with modularity of 0.619, while ten modules for EABCA with modularity of 0.621. ForceAtlas2 was applied in the layouts as shown in Figure 8 & Figure 9. The modules' details are shown in Table 10.

Table 10. Modules identified by ABCA and EABCA in Network Analysis

Module	*ABCA*	*EABCA*	*MeSH*	*Keywords*
A	Multiple Myeloma		Multiple Myeloma / Lenalidomide / Gene Rearrangement	lenalidomide / relapsedrefractory / ixazomib
A-1		Multiple Myeloma and Therapy	Multiple Myeloma / Lenalidomide / Combined Modality Therapy	myeloma / relapsedrefractory / bortezomib
A-2		Multiple Myeloma and Genomics	Multiple Myeloma / DNA Copy Number Variations / Genomics	myeloma / genomic / genetic
B	Immunotherapy in Melanoma and Esophageal Cancer		Melanoma / Nivolumab / Stomach Neoplasms / Esophagogastric Junction / Programmed Cell Death 1 Receptor	melanoma / gastric / checkpoint
B-1		Melanoma and Immunotherapy	Antineoplastic Agents, Immunological / Melanoma / Immunotherapy / CTLA-4 Antigen	melanoma / checkpoint / immunotherapy
B-2		Esophageal & Gastric Cancer	Stomach Neoplasms / Esophagogastric Junction / Esophageal Neoplasms	gastric / junction / attraction-2
C	Myeloid Sarcoma		Hematopoietic Stem Cell Transplantation / Leukemia, Myeloid, Acute / Myeloproliferative Disorders / Primary Myelofibrosis	myeloid / myelofibrosis / aml
C-1		Stem Cell Transplantation	Graft vs Host Disease / Hematopoietic Stem Cell Transplantation / Transplantation, Homologous	transplantation / ebmt / allogeneic
C-2		Acute Myeloid Leukemia and Therapy	Leukemia, Myeloid, Acute / Myelodysplastic Syndromes / Remission Induction	myeloid / myelodysplastic / aml
C-3		Myeloproliferative Disorders and Therapy	Primary Myelofibrosis / Myeloproliferative Disorders / Thrombocythemia, Essential	myelofibrosis / myeloid / ruxolitinib
D	RNA Interference	RNA Interference	MicroRNAs / RNA, Circular / RNA Interference	rna / pathway / lncrna
E	Lymphoma and therapy	Lymphoma and therapy	Lymphoma, Large B-Cell, Diffuse / Lymphoma / Lymphoma, Non-Hodgkin	lymphoma / b-cell / dlbcl
F		Colorectal Neoplasms and Proteomics	Colorectal Neoplasms / Protein Kinase Inhibitors / Proteomics	colorectal / muc1-c / mastl

From the macro perspective of the layout, the division among EABCA's modules is clearer, and there are also more subdivided modules, which appear more compact. In ABCA, the modules C, A & D are in a triangular relationship that closes to each other. While in EABCA, there exists a gap from C-1, C-2 & C-3 to A-1 & A-2, and even a far distance to D.

From the meso perspective of the module, some large modules are subdivided in EABCA. Figure 10 shows how these nodes migrated from ABCA to EABCA. Module C# Myeloid Sarcoma is subdivided into three closely related modules C-1, C-2, and C-3. Module B is divided into B-1 that focuses on immunotherapy in melanoma, and B-2 on esophageal & gastric cancer. Module A# Multiple Myeloma is divided into A-1 that focuses on therapy and

A-2 on genomics. We also find that B-2 includes authors transferred from D and engaged in gastric cancer research, represented by To, Ka Fai. Besides, a small immature module F in EABCA involving colorectal neoplasms is composed of nodes moved from B, C, and D.

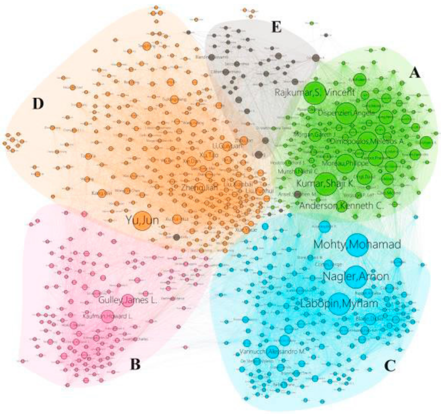


Figure 8. Network layout of ABCA

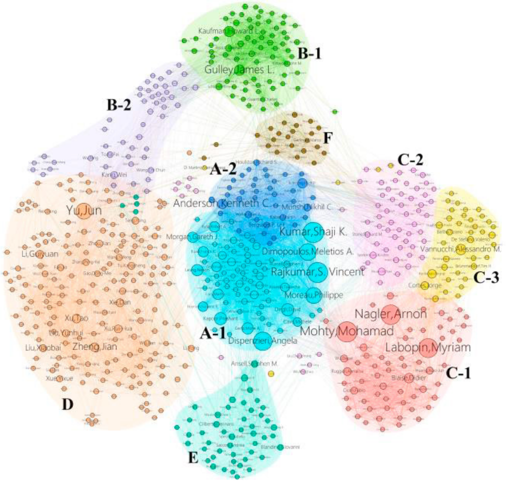


Figure 9. Network layout of EABCA

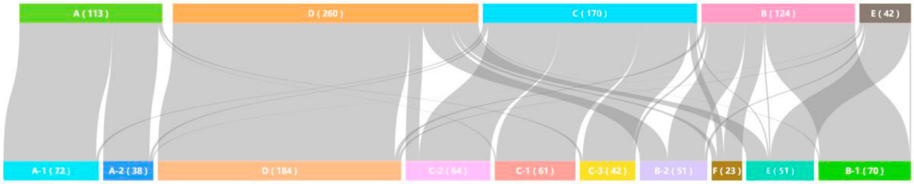


Figure 10. How do the nodes in modules of ABCA migrate to modules of EABCA

Table 11. Example of the different citation context of single reference

Reference PMID	*CON_SIM*	*BC Authors*	*Citation Context in Citing Articles*	*Citing Title*
24836762	0.289	Mohty, Mohamad	['... Several studies have demonstrated feasibility of azacytidine during the post-transplant period, and the correlation of the expansion determined cytotoxic T cell subsets against tumour antigens with a lower relapse incidence [37]…']	['Allogeneic stem cell transplantation in adult patients with acute myeloid leukaemia and 17p abnormalities in first complete remission...']
		Anderson, Kenneth C.	['...These need not be myeloma-specific panels, as many vendors provide kits or services for TP53[25]...']	['A high-risk, Double-Hit, group of newly diagnosed myeloma identified by genomic analysis']

From the micro perspective, the above phenomenon, including the gap, subdivision, and migration, could be explained from key nodes. For example, in EABCA, the edge weight between the representative author Mohty, Mohamad of C-1 and Anderson, Kenneth of A-2, shrinks from 15 in ABCA to 1.7407; hence, the distance between them is greatly extended. Indeed, there are apparent differences in their research interests, whose BC is limited to transplantation therapy in Myeloma, or some commonly involved substance, genes, etc. As shown in Table 11, although they both cite one paper, the contexts greatly vary. The strength correction also occurs between other key nodes, such as Mohty, Mohamad and Kumar, Shaji

K. of A-1 (drops from 10 to 2.816); Dispenzieri, Angela of A-1 and Li, Guiyuan of D (from 2 to 0.527); Li, Guiyuan and Kang, Wei of B-2 (from 5 to 0.699); Kaufman, Howard L. of B-1 and Doi, Toshihiko of B-2 (from 10 to 2.179); Gulley, James L. of B-2 and Yu, Jun of D (from 4 to 1.017); Mohty, Mohamad and Stone, Richard M. of C-2 (from 10 to 3.163), and Cortes, Jorge of C-3 (from 19 to 5.019); Li, Guiyuan and Blandino, Giovanni of E (from 5 to 0.558), Cerham, James R. of E and Dimopoulos, Meletios A. of A-1 (from 6 to 0.387), etc..

Discussion and Conclusion

In summary, the results show EABCA has the following advantages over ABCA:

(1) EABCA can more accurately reveal the association between authors, which is manifested as follows: The matrix generated by it obtain clustering results with higher performance, the edge weight between key nodes is verified to be in line with reality better. That is mainly because EABC relevance is calculated based on the full-text level data source and comprehensively considers multiple dimensions instead of solely relying on explicit counts. This kind of richer scale may also help us reason what happened in the black box of BC.

(2) EABCA can find a more detailed intellectual structure, especially among some research interest groups that are not clearly distinguished, as shown in factor and network analysis. It is known that scientific literature usually needs to intensively cite relevant works, which leads to rather dense BC relations. The fact makes it difficult to subdivide emerging sub-fields from other close or parent fields when the strength solely relies on the count, while EABCA's particular strength is competent. This is because there exist differences in terms, contexts, etc., among research topics; hence the citations may appear with different syntactic and semantic characteristics when researchers cite literature with various degrees of relevance. On this basis, detailed research topics and communities are easier to distinguish in EABCA.

This study shows that expanding the depth of data in ABCA can improve its analysis effect, revealing that it is feasible and meaningful to apply CCA further in promoting precision bibliometrics. Also, this effect improvement may positively impact a broader range of high-level applications, i.e., academic cooperation prediction and recommendation.

The study's main limitation is the incompleteness of full-text resources available; therefore, our empirical output results only have practical value in a specific range. Besides, we believe EABCA still has room for further refinement. For example, due to the inconsistent section titles in BC articles, we applied an alternative location tagging method to overcome large-scale missing location data; the better solution is to map a standardized structure. Moreover, the implications of the full-author pattern and the discipline specificity on EABCA are still not explicit, requiring further exploration and a wider range of empirical studies.

Acknowledgments

We thank Prof. Ronald Rousseau for his valuable suggestions, Dr. Yining Zhang of The First Hospital of China Medical University, Dr. Xiaoxue Zhang of Liaoning Cancer Hospital & Institute for their help in identifying and labeling topics. This work is supported by LIS Outstanding Talents Introducing Program, Bureau of Development and Planning of CAS [2018] No. 12.

References

An, J., Kim, N., Kan, M.-Y., Chandrasekaran, M. K., & Song, M. (2017). Exploring characteristics of highly cited authors according to citation location and content. *Journal of the Association for Information Science and Technology*, 68(8), 1975-1988.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.

Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), 45-80.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, 7(3), 583-592.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, 65(9), 1820-1833.

Gipp, B., & Beel, J. (2009). Citation proximity analysis (CPA)—A new approach for identifying related work based on co-citation analysis. In B. Larsen, & J. Leta (Eds.),*Proceedings of the 12th international conference on scientometrics and informetrics (ISSI'09)* (pp. 571-575). Rio de Janeiro (Brazil): International Society for Scientometrics and Informetrics.

Glänzel, W., & Thijs, B. (2011). Using 'core documents' for the representation of clusters and topics. *Scientometrics*, 88(1), 297-309.

Habib, R., & Afzal, M. T. (2019). Sections-based bibliographic coupling for research paper recommendation. *Scientometrics*, 119(2), 643-656.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). Multivariate data analysis (7th ed.).

Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics*, 8(1), 197-211.

Kessler, M. M. (1963). Bibliograpic coupling between scientific papers. *American Documentation*, 14(1), 10-25.

Kim, H. J., Jeong, Y. K., & Song, M. (2016). Content- and proximity-based author co-citation analysis using citation sentences. *Journal of Informetrics*, 10(4), 954-966.

Kumar, V., Sendhilkumar, S., & Mahalakshmi, G. S. (2017). Author Similarity Identification Using Citation Context and Proximity. *2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM).*

Liu, S., & Chen, C. (2011). The proximity of co-citation. *Scientometrics*, 91(2), 495-511.

Liu, X., Zhang, J., & Guo, C. (2013). Full-text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology*, 64(9), 1852-1863.

Leydesdorff, L. (2011). BibCoupl.exe for Bibliographic Coupling among Authors. Retrieved January 21, 2021 from https://www.leydesdorff.net/software/bibcoupl/index.htm.

Ma, R. (2012). Author bibliographic coupling analysis: A test based on a Chinese academic database. *Journal of Informetrics*, 6(4), 532-542.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In 1st International Conference on Learning Representations, ICLR 2013, Scottsdale.

Rousseau, R. (2010). Bibliographic coupling and co-citation as dual notions. The Janus faced scholar. *A Festschrift in honour of Peter Ingwersen*, 173-183.

Ružička, M. (1958). Anwendung mathematisch-statistischer Methoden in der Geobotanik (Synthetische Bearbeitung von Aufnahmen). Biológia, Bratislava, 13: 647-661.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, 49(4), 327-355.

Zhang, S., Xu, Y., & Zhang, W. (2019). Clustering Scientific Document Based on an Extended Citation Model. *IEEE Access*, 7, 57037-57046.

Zhao, D., & Strotmann, A. (2008a). Evolution of research activities and intellectual influences in information science 1996-2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science and Technology*, 59(13), 2070-2086.

Zhao, D., & Strotmann, A. (2008b). All-author vs. first-author co-citation analysis of the Information Science field using Scopus. *Proceedings of the American Society for Information Science and Technology*, 44(1), 1-12.

Zhao, D., & Strotmann, A. (2014). The knowledge base and research front of information science 2006-2010: An author cocitation and bibliographic coupling analysis. Journal of the Association for Information Science and Technology, 65(5), 995-1006.

Applying Bibliometrics Methods to Understanding Knowledge Domains on Wikipedia

Dangzhi Zhao[1] and Andreas Strotmann[2]

[1] *dzhao@ualberta.ca*
School of Library and Information Studies, University of Alberta, Edmonton (Canada)

[2] *andreas.strotmann@gmail.com*
ScienceXplore, F.-G.-Keller-Str. 10, D-01814 Bad Schandau (Germany)

Abstract

Bibliometrics is effective in studying scholarly communication patterns and mapping domains of knowledge published in scholarly documents. The present study explores what we may learn from applying bibliometric methods to studying domains of knowledge represented in Wikipedia, a gigantic encyclopaedia created and maintained collectively by volunteers world-wide. In particular, we adapted and applied author bibliographic coupling analysis to examine one of the many topic areas on Wikipedia: Traditional Chinese Medicine (TCM). We found that seven dimensions of TCM are represented on WP: Acupuncture and related practices, Falun Gong, Herbal Medicine, TCM and its common concepts, Qigong, Chinese Martial Arts, and non-herbal medicine. The first three dimensions attracted the most Wikipedia contributors to TCM. Acupuncture and Qigong have the most connections to the TCM field, and also serve as bridges for other topics to connect to the TCM filed. Herbal medicine is linked to the field through the generalists who made low contributions to many TCM topic areas. Non-herbal Medicine is isolated from the rest of the field. It appears that specific topics are represented well on WP but their conceptual connections are not, especially between a general topic (e.g., TCM and its common concepts) and its sub-topics (e.g., Qigong, Taichi).

Introduction, background and related studies

Wikipedia (WP) is a system unique in the history of civilization (Simonite, 2013). Both benefits and challenges of the WP system have been widely debated in academia, law, business, and other sectors of society.

WP started in 2001 with the lofty goal of compiling all human knowledge. It grew quickly into the largest encyclopaedia in the world (Burke & Kraut, 2008). As of January 2021, WP has over 40.7 million registered and uncounted unregistered volunteer editors and over 6.2 million articles in the English version alone (WP:Statistics). Many WP articles were of a quality comparable with corresponding ones in Encyclopaedia Britannica (Giles, 2005). WP's success also made it play a symbolic role in highlighting the potential for voluntary peer-production to generate valuable collections of information. Hansen, et al., (2009) even contend that WP "approximates features of the ideal speech situation articulated by Habermas" in his Theory of Communicative Actions (Habermas, 1984). WP has now become a one-stop shop for information on pretty much any topic. As WP articles often intentionally show up at the top of Google search result lists and are now promoted as fact check sources on Facebook, to name just two prominent examples, WP is clearly a primary information source that people see, and in many cases even the only source that people use.

However, WP's success was to many a surprise. Among the fundamental problems of the WP system that have been criticized are unpredictable motivations (and competences) of editors and an emphasis on consensus rather than authority (Denning et al., 2005). WP evolves without supervision by certified subject experts or authorities, and its largely anonymous volunteer editors are left both to select, write about and organize the topics it includes and to define, interpret and implement its policies and resolve conflicts on their own. WP editors may be knowledge domain experts or an elementary school student, and "may be altruists, political or commercial opportunists, practical jokers, or even vandals" (Denning et al., 2005, p. 152). Inaccuracy or errors may exist due to lack of supervision by certified subject experts. Bias can be introduced and maintained as long as a group of editors with that bias manage to dominate the discussion and force it to a "consensus." Mechanisms used in traditional systems to ensure quality and avoid abuse of power are normally based on true identity along with social expectations, norms, and status positions, and thus cannot work for WP (Arazy et al., 2011; Ransbotham and Kane, 2011).

The incident when WP rejected an entry about Donna Strickland, a Canadian female winner of the 2018 Nobel Prize in Physics, half year before the announcement of the prize is just one of the many examples of problems in WP's topic selection policy and practices (The Guardian, 2018). WP categories have been found to be "very loosely structured" (Perez, 2021) and therefore there have been many studies on automatic categorization of WP articles (e.g., Refaei et al, 2018; Perez, 2021; Simone Paolo Ponzetto and Strube, 2007; Strube, and Simone Paolo Ponzetto, 2006; Gantner & Schmidt-Thieme, 2009). Despite these problems, as the largest open encyclopaedia, the WP's content (articles and categories) "has been used extensively for tasks like entity disambiguation or semantic similarity estimation" to enhance AI-based knowledge representation and organization (Heist and Paulheim, 2019).

Given the prominent position WP has for both information users and for researchers on knowledge organization and information retrieval, it is important to study how well knowledge is represented and structured on WP. The present study applies one of the effective bibliometric methods for mapping knowledge domains represented in published scholarly documents, i.e., author bibliographic coupling analysis (Zhao and strotmann, 2008), to examine one of the many topic areas on Wikipedia: Traditional Chinese Medicine (TCM). We chose the TCM topic area for this study because it has been reported to be one of the most controversial topic areas on WP in which WP's problems in topic selection and treatment may be more pronounced (Koppelman, 2017; McLuhan, 2013).

Methodology

Author bibliographic coupling analysis (ABCA)

Verifiability is one of the two fundamental principles among all the detailed policies and guidelines for contributing and for resolving conflicts that WP has developed, the other being Neutral Point of View (NPOV). These principles intend to ensure all important viewpoints are represented fairly and supported by trustworthy published resources (cited references).

Bibliographic coupling method uses the number of cited references shared by two publications to measure how closely these two articles are related in terms of topics or methodological approaches. The degree to which they share references has also been used to measure topic similarity between WP articles (Das et al., 2016).

Author bibliographic coupling analysis (ABCA) uses author instead of article as the unit for analysis. ABCA has been found to have a number of advantages compared to article-based analysis (Zhao and Strotmann, 2008). One of these advantages is the wider granularity and more data an author represents that can smooth out impact of outliers on the analysis. Authors represent schools of thought whereas articles represent individual pieces of evidence for or findings about concepts, theories or methods. Individuals tend to develop a set of information sources that they prefer to consult when they contribute to knowledge production (e.g., writing scholarly or WP articles) or deal with work or life problems (White, 2001). The more information sources are shared by two individuals the more closely their interests and beliefs may be related.

Data collection and analysis

We adapted and followed the well-established procedures and techniques for ABCA. Instead of using a citation database, we used the English version of WP as data source, and developed computer programs to collect and analyze data from WP.

We identified articles on Traditional Chinese Medicine (TCM) from WP by starting with the articles on TCM proper and with articles under the sub-categories and sub-sub-categories of TCM. We downloaded all these articles from WP in late 2019, including the entire editing and discussion history of each article.

Following ABCA techniques, we chose the top 500 editors who have contributed the most to the downloaded articles to represent this topic area. A matrix of shared reference scores was produced for all these editors. Specifically, if editors A and B have contributed substantially to article sets S1 and S2 respectively, and n different information sources were cited in both S1 and S2, n would be the shared reference score for A and B. We deleted those editors whose vectors contain only zeros, i.e., those who do not share cited references with any other editors in the set, which resulted in a 380x380 matrix. We left the diagonal cells empty in this matrix as in previous studies. They were treated as missing data and replaced with the mean in the Factor Analysis routine in SPSS that we used to explore the underlying structure of the interrelationships between these editors.

Factors were extracted by Principal Component Analysis (PCA). The number of factors extracted was determined based on an examination of the Scree plot, total variance explained, and correlation residuals – the differences between observed correlations and correlations implied by the factor model (Hair, et al., 1998). This resulted in an 18-factor model that explains 57.5% of the total variance, and the differences between observed and implied correlations are smaller than 0.05 for the most part (92%).

An oblique rotation was applied resulting in a pattern matrix and a structure matrix. We use the highest loading of a factor in the pattern matrix to indicate its distinctiveness. The size or prominence of a factor is indicated by the number of editors who load primarily on this factor in the pattern matrix. A Component Correlation Matrix showing how closely factors are related to each other was also produced by the Factor Analysis routine.

Factors are labeled upon manual examination of articles that authors who load primarily in each factor have significantly contributed references to. A factor is labeled as undefined (UD) if all loadings in this factor are lower than 0.7, although an attempt may still be made at interpreting it.

Results and discussion

Table 1 shows the topics identified and their distinctiveness and prominence indicated by the highest loading and the number of editors loading primarily and significantly on a factor respectively. Table 2 is the Component Correlation Matrix showing how closely these topics are related to each other. Correlations that are higher than 0.2 are highlighted and are considered as indicating a close connection in the discussions below.

Table 1. Topics and their prominence (Size) and distinctiveness (Highest loading)

Factor	*Topical Label*	*Size*	*Highest_loading*
1	Acupuncture	84	0.998
2	Falun Gong - destructive	35	0.998
3	Falun Gong - constructive	35	1.011
4	Generalists	20	0.945
5	Chinese martial arts	16	0.952
6	TCM and its Common concepts	19	0.989
7	Fungi used in TCM	19	0.894

Table 1. (*cont'd*) Topics and their prominence (Size) and distinctiveness (Highest loading)

8	Qigong	14	0.982
9	Chinese Massages	17	0.805
10	Goji	21	0.861
11	UD	7	0.661
12	Tai chi	19	0.852
13	Medicinal plants	20	0.892
14	UD	11	0.681
15	Psychoactive plants	11	0.832
16	UD (Flowering plants)	14	0.674
17	Plants used as both spice and medicine	12	0.802
18	Tiger bone wine	6	0.944

Table 2. Component Correlation Matrix

Factor	1	2	3	4	5	6	7	8	9
1	1.000								
2	0.016	1.000							
3	-0.006	0.621	1.000						
4	0.210	0.047	0.040	1.000					
5	0.138	0.133	0.115	0.253	1.000				
6	0.380	0.013	0.035	0.125	0.008	1.000			
7	0.176	0.037	0.043	0.399	0.133	0.109	1.000		
8	0.289	0.147	0.217	0.197	0.216	0.197	0.194	1.000	
9	0.256	-0.051	0.003	0.077	0.022	0.175	0.030	0.301	1.000
10	0.012	-0.031	-0.037	0.203	-0.019	0.150	0.171	0.240	0.059
11	0.016	0.090	0.034	-0.005	0.053	0.155	-0.075	-0.019	-0.172
12	0.079	0.069	0.046	0.272	0.323	0.048	0.196	0.072	-0.038
13	0.016	0.079	0.040	0.320	0.026	0.064	0.272	0.042	-0.019
14	0.243	0.037	0.070	0.279	0.151	0.195	0.171	0.235	0.191
15	0.050	0.025	0.005	0.101	0.010	0.107	0.197	0.126	0.126
16	0.120	0.120	0.147	0.109	0.163	0.008	0.218	0.120	-0.033
17	0.016	-0.026	-0.017	0.133	-0.040	0.038	0.210	0.011	0.037
18	-0.010	-0.006	-0.006	-0.004	0.004	-0.020	0.010	-0.007	-0.017

Component	10	11	12	13	14	15	16	17	18
10	1.000								
11	-0.011	1.000							
12	0.021	0.142	1.000						
13	0.133	0.172	0.142	1.000					
14	0.155	-0.221	0.102	0.029	1.000				
15	0.131	-0.087	0.089	0.117	0.106	1.000			
16	0.045	-0.152	0.103	0.002	0.169	0.222	1.000		
17	0.056	0.043	0.101	0.073	0.029	-0.023	-0.038	1.000	
18	-0.012	-0.029	-0.002	-0.018	-0.001	-0.002	0.028	0.004	1.000

Structure of the TCM field on WP

Identified topics in Table 1 can be grouped into seven dimensions: TCM and its common concepts (F6), Acupuncture and related practices (F1, F9), Qigong (F8), Falun Gong (F2, F3), Herbal Medicine (F7, F10, F13, F15, F16, F17), Chinese Martial Arts (F5, F12), and non-herbal medicine (F18).

Acupuncture and related practices, Falun Gong, and Herbal Medicine are the most prominent topic areas as indicated by the large numbers of WP editors associated with them primarily. Among the identified topics, Acupuncture and Qigong have the most connections to the TCM field, with 5 and 6 highlighted correlations in Table 2 respectively. They also serve as bridges for other topics to connect to the field: "TCM and its common concepts" connects with Acupuncture which connects to the rest of the field; Falun Gong and Chinese Martial Arts only connect to the rest of the filed through Qigong. The strongest connections (i.e., correlation above 3) are between Acupuncture and TCM concepts, and between Qigong and Chinese martial arts. It is somewhat surprising to see that TCM concepts are perceived on WP as only closely related to a single topic area: Acupuncture, which suggests that TCM concepts have not been applied to the discussions of the other conceptually related TCM topics such as Qigong and Taichi.

Herbal medicine has many different types such as those that are also used as spices (e.g., Cinnamomum cassia) and those used as psychoactive drugs (e.g., Cannabis). This topic area is linked to the TCM field through the generalists who made low contributions to all TCM topics except for Falun Gong and Tiger Bone Wine. In contrast, only one non-herbal medicine topic (Tiger bone wine) stands out as a separate factor, and this very small area is largely isolated from the rest of the TCM field indicated by the very low correlations with all other factors.

Topics of the TCM field on WP

The most prominent topic is Acupuncture (F1), which aligns well with the fact that among TCM theories, techniques and practices, acupuncture is the most recognized, accepted and practiced in the Western world. In many western countries, acupuncture therapists are trained and certified, and acupuncture treatments are covered by most health insurance plans while other TCM practices such as herbal medicine are not. Qigong (F8) and Chinese massages (F9; cupping therapy, Guasha, etc.) are practices closely related to acupuncture (Table 2), but are much less recognized as seen from their much smaller sizes (Table 1).

The second most prominent topic is Falun Gong, which has a size only slightly smaller than Acupuncture when considering both constructive (F3) and destructive (F2) contributions to the topic. Falun Gong has been highly controversial and attracted attention from many. Editors who belong to the constructive group tend to have broad interest in TCM beyond Falun Gong and added or modified significantly more references than they removed on Falun Gong. In contrast, editors who belong to the destructive group tend to be highly focused on Falun Gong and removed significantly more references than they added or modified.

Falun Gong is largely separated from the rest of the TCM field as indicated by the mostly very low correlations with all the other topics. The only relatively high correlation the two Falun Gong factors have is with Qigong, which recognizes their conceptual connection. We can speculate on reasons for this separation. Although there are controversies over several TCM topics, such as acupuncture and Chinese herbal medicine, those are mostly about whether they are scientific and effective. Controversies around Falun Gong, however, are highly political. Essentially, editors who contributed to Falun Gong related WP articles may not be interested in TCM per say but in politics whereas editors who contributed to other TCM topics were drawn to the medicine aspects of the topics.

It is interesting to see that sub-areas of herbal medicine are not interrelated to each other but form a weak partial linear chain of links: Fungi used in TCM (F7) is related to Plants used as both spice and medicine (F17) and to the topic that appears to be flowering plants such as rose and magnolia (F16) which is then related to psychoactive plants (F15). Goji (F10) and the generic topic area on medicinal plants (F13) are not related to any other sub-areas of herbal medicine. Similar to the earlier observation that TCM concepts are perceived on WP as only closely related to a single topic area: Acupuncture, it is somewhat surprising to see that the general topic on medicinal plants is not closely related to any of the specific types of medical plants.

Several sub-areas of herbal medicine, including the generic topic area on medicinal plants (F13), Plants used as both spices and medicine (F15), and Psychoactive plants (F17), have a relatively high correlation with only one of the other TCM topics. The topic "Plants used as both spice and medicine" (F17) is largely separated from the rest of the TCM field, with mostly very low correlations with all the other topics. Contributors to this area who are only interested in the plants as spice may have separated this area out from the rest of TCM.

Taichi (F12) is correctly recognized on WP as a form of Chinese martial arts as indicated by its relatively strong correlation with Chinese martial arts (F5). It is interesting to see that Qigong (F8) is perceived on WP as more closely related to Chinese massages (F9) than to Chinese martial arts although it is often mentioned together with both in the literature and movies.

Conclusions

The present study applied one of the effective bibliometric methods for mapping knowledge domains represented in published scholarly documents, i.e., author bibliographic coupling analysis, to examine one of the many topic areas on Wikipedia: Traditional Chinese Medicine (TCM). It is interesting to explore what we may learn from bibliometric studies of domains of knowledge represented in Wikipedia, a gigantic encyclopaedia created and maintained collectively by volunteers world-wide.

We found that seven TCM topic areas are represented on Wikipedia, among which Acupuncture related practices, Falun Gong, and Herbal Medicine attracted the most of significant contributors to TCM. Acupuncture and Qigong have the most connections to the TCM knowledge domain, and also serve as bridges for other topics to connect to the domain. Herbal medicine is weakly linked to and non-herbal medicine is isolated from the rest of the TCM knowledge domain.

It appears that specific topics are represented well on WP but their conceptual connections are not, especially those between a general topic (e.g., TCM and its common concepts) and its sub-topics (e.g., Qigong, Taichi).

The present study shows that ABCA is effective for mapping knowledge domains represented on Wikipedia.

Acknowledgments

This study was supported by an Insight Development Grant from the Social Sciences and Humanities Research Council of Canada (SSHRC). We would like to thank our research assistant Xinru Zhang for writing some of the data collection and analysis programs for this study.

References

Arazy, O., Nov, O., Patterson, R., & Yeo, L. (2011). Community-Based Collaboration in Wikipedia: The Effects of Group Composition and Task Conflict on Information Quality. *Journal of Management Information Systems,* 21(4), 71–98.

Burke, M., & Kraut, R. (2008). Mopping up: modeling Wikipedia promotion decisions. *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (pp. 27-36).

Das, S., Lavoie, A., & Magdon-Ismail, M. (2016). Manipulation among the arbiters of collective intelligence: How Wikipedia administrators mold public opinion. *ACM Transactions on the Web*, 10 (4).

Denning, P., Horning, J., Parnas, D. and Weinstein, L. (2005). Wikipedia risks. *Communications of the ACM*, 48(12): 152.

Gantner, Z., and Schmidt-Thieme, L. (2009). Automatic Content-based Categorization of Wikipedia Articles. *People's Web '09: Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, Pages 32–37. Retrieved Jan 19, 2021 from https://www.aclweb.org/anthology/W09-3305.pdf

Giles, G. (2005). Internet Encyclopedias Go Head to Head. *Nature*, 438(7070), 900-901.

Habermas, J. (1984). *The theory of communicative action: Reason and the rationalization of society*. Boston: Beacon Press.

Hansen, S., Berente, N., & Lyytinen, K. (2009). Wikipedia, Critical Social Theory, and the Possibility of Rational Discourse. *The Information Society – An International Journal*, 25(1), 38-59.

Heist N., Paulheim H. (2019) Uncovering the Semantics of Wikipedia Categories. In: Ghidini C. et al. (eds) *The Semantic Web – ISWC 2019. ISWC 2019. Lecture Notes in Computer Science*, 11778. Springer, Cham. https://doi.org/10.1007/978-3-030-30793-6_13.

Koppelman, M.H. (2017). WikiTweaks: The Encyclopaedia that Anyone (Who is a Skeptic) Can Edit. *Journal of Chinese Medicine*. Feb2017, Issue 113, p35-40.

McLuhan, R. (2013). *Guerrilla Skeptics*. Retrieved Dec. 18, 2018 from: https://monkeywah.typepad.com/paranormalia/2013/03/guerrilla-skeptics.html.

Perez, B., West, A.G., Feo, C., and Lee, I. (2021). *WikiCat: A graph-based algorithm for categorizing Wikipedia articles*. Retrieved Jan 19, 2021 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.641.5224&rep=rep1&type=pdf.

Refaei1, N., Hemayed, E.E., and Mansour, R. (2018). WikiAutoCat: Information Retrieval System for Automatic Categorization of Wikipedia Articles. *Arabian Journal for Science and Engineering 43*, 8095–8109. https://doi.org/10.1007/s13369-018-3244-9. Retrieved Jan 19, 2021 from https://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=e15723ec-a8e9-4c6b-b7f4-da9e067ea2ad%40sdc-v-sessmgr03

Ransbotham, S., & Kane, G.C. (2011). Membership Turnover and Collaboration Success in Online Communities: Explaining Rises and Falls from Grace in Wikipedia. *MIS Quarterly* 35(3), 613–627.

Simone Paolo Ponzetto and Strube, M. (2007). Deriving a large scale taxonomy from wikipedia. *Proceedings of the 22nd national conference on Artificial intelligence*, 2, 1440–1445.

Simonite, T. (2013). The Decline of Wikipedia. *MIT Technology Review*. Retrieved Nov. 12, 2018 from: https://www.technologyreview.com/s/520446/the-decline-of-wikipedia/.

Strube, M., and Simone Paolo Ponzetto. (2006). Wikirelate! computing semantic relatedness using wikipedia. *Proceedings of the 21st national conference on Artificial intelligence,* 2, 1419–1424.

The Guardian (2018). *Female Nobel prize winner deemed not important enough for Wikipedia entry*. Retrieved Jan 19, 2021 from https://www.theguardian.com/science/2018/oct/03/donna-strickland-nobel-physics-prize-wikipedia-denied

Zhao, D., & Strotmann, A. (2008). "Evolution of research activities and intellectual influences in Information Science 1996-2005: Introducing author bibliographic coupling analysis". *Journal of The American Society for Information Science and Technology,* 59(13), 2070-2086.

International Migration in Academia and Citation Performance: An Analysis of German-Affiliated Researchers by Gender and Discipline Using Scopus Publications 1996-2020

Xinyi Zhao[1], Samin Aref[1], Emilio Zagheni[1] and Guy Stecklov[2]

[1] *{zhao, aref, zagheni}@demogr.mpg.de*
Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research,
Konrad-Zuse-Str. 1, 18057 Rostock (Germany)

[2] *guy.stecklov@ubc.ca*
Department of Sociology, University of British Columbia, 6303 NW Marine Drive, Vancouver, BC (Canada)

Abstract

Germany has become a major country of immigration, as well as a research powerhouse in Europe. As Germany spends a higher fraction of its GDP on research and development than most countries with advanced economies, there is an expectation that Germany should be able to attract and retain international scholars who have high citation performance. Using an exhaustive set of over eight million Scopus publications, we analyze the trends in international migration to and from Germany among published researchers over the past 24 years. We assess changes in institutional affiliations for over one million researchers who have published with a German affiliation address at some point during the 1996-2020 period. We show that while Germany has been highly integrated into the global movement of researchers, with particularly strong ties to the US, the UK, and Switzerland, the country has been sending more published researchers abroad than it has attracted. While the balance has been largely negative over time, analyses disaggregated by gender, citation performance, and field of research show that compositional differences in migrant flows may help to alleviate persistent gender inequalities in selected fields.

Introduction

In the current era of knowledge-based economies, highly skilled people are the most mobile population group worldwide (Schiller & Cordes, 2016). In OECD countries, highly qualified individuals, including researchers, make up one-third of the immigrant population (Docquier & Marfouk, 2004; Schiller & Cordes, 2016). The international mobility of researchers facilitates the exchange of knowledge, ideas, and skills, and thus contributes to the dynamic development of the global knowledge production system (Bauder, 2015; Franzoni et al., 2015; Netz & Jaksztat, 2017). Countries see researchers as prized groups because of their high levels of human capital and potential for fueling innovation and economic growth. Gaining a better understanding of this kind of mobility is crucial for evaluating scientific research at the national level, and for informing policy on science and mobility in academia (Netz et al., 2020). As Germany is a science powerhouse that hosts more non-EU researchers than any other EU country (Jöns, 2009; IDEA Consult, 2013; Lörz et al., 2016; Guthrie et al., 2017; Aman, 2018), Germany is expected to be able to attract and retain international researchers with high citation performance. A lack of fine-grained data on the international migration of German-affiliated researchers and their citation performance makes it difficult to understand the in- and out-flow patterns and costs associated with this kind of mobility. This study develops individual-level migration data in order to analyze the mobility of researchers to and from Germany, while taking into account the citation performance of internationally mobile scholars across disciplines and genders.

Germany's relatively liberal immigration laws and its powerful position as the world's fourth-largest economy (Schiller & Cordes, 2016) are key factors in explaining why about one-fifth of the top-level researchers in Germany in 2004 were foreign-born (Ioannidis, 2004). Over the past three decades, German universities and research institutions have attracted large numbers of researchers from other countries, especially from other European countries (Giousmpasoglou

& Koniordos, 2017). In addition, Germany spends over 3.1% of its gross domestic product (GDP) on research and development (R&D), which is more than the R&D spending by GDP in most OECD countries, including the United States (US) (2.8%) and the United Kingdom (UK) (1.7%). It has been suggested that highly cited researchers tend to migrate to countries with higher R&D spending (Hunter et al., 2009). However, when the two migration directions are considered, Germany appears to have suffered a net loss of 28% due to the global migration of researchers (Schiller & Cordes, 2016; OECD, 2008). More specifically, Germany is experiencing a "brain drain" in certain specializations, including medical research (Giousmpasoglou & Koniordos, 2017). Although existing studies on the German science system have explored some aspects of academic migration (Netz & Finger, 2016; Aman, 2016; Parey et al., 2017; Netz & Jaksztat, 2017; Netz & Grüttner, 2020), there has been no systematic analysis of international academic migration for Germany that has taken the citation performance of researchers into account, and that has covered all fields of scholarship. This study provides a much-needed baseline for the development of future policies that can help Germany succeed in attracting and retaining qualified researchers from overseas, facilitating the circulation of brainpower, and advancing the performance of the German science system.

It is generally recognized that international academic mobility has a positive influence on the global scientific system. By promoting knowledge production and diffusion between countries, migration among researchers enhances the performance of global science production (Giousmpasoglou & Koniordos, 2017; Guthrie et al., 2017; Netz et al., 2020). From a micro perspective, there is strong evidence that mobile researchers outperform non-mobile researchers (Dubois et al., 2014; Franzoni et al., 2014; Moed & Halevi, 2014; Guthrie et al., 2017). Findings from the MORE survey have shown that mobility leads to increased outputs for researchers in both academia and industry. Notably, there is evidence that the output effects are higher among researchers moving from the EU to the US than among those moving in the opposite direction (Børing et al., 2015). Gibson and McKenzie found that emigrant researchers have much greater research outputs and impacts, as measured by total citation and *h*-index, than researchers who stay in their respective origin countries (Gibson & McKenzie, 2014). A comprehensive evaluation of the interplay between migration and citation performance that takes individual all disciplines into account can deepen our understanding of this crucial topic.

Qualitative interviews, surveys, bibliometric data, and data from curricula vitae are among the most common data sources for studying the migration of researchers (Netz & Jaksztat, 2017; Netz & Grüttner, 2020). The idea of using the historical records of researchers to follow their geographical movements can be traced back to a study by Rosenfeld & Jones (1987) on the movements of psychologists in the US that used a sample from the biographies of members of the American Psychology Association. The digital revolution and the advent of digitized sources of bibliometric information enables us to expand this simple idea to cover a large number of data points with a flexible level of granularity that is suitable to our research objectives. Previous studies that utilized similar applications of bibliometric data have mapped academic mobility among countries, and have shed light on the causes and consequences of academic mobility (Moed & Halevi, 2014). Recent methodological innovations for re-purposing bibliometric data to study internal migration within country boundaries have made the process of inferring migration events from affiliation addresses more reliable (Miranda-González et al., 2020).

This paper relies on large-scale digitized bibliometric data from *Scopus* to document and analyze international migration to and from Germany among researchers over the past 24 years. The cleaned and pre-processed bibliometric data provide a unique perspective on the international migration patterns and geographical trajectories of mobile researchers. The results help to clarify the position of Germany in the global science system. In addition, the analysis evaluates the interplay between migration and citation performance across different disciplines.

Thus, this analysis adds a demographic dimension to the science of science literature by providing several in-depth statistics related to academic mobility.

Materials and methods

Scopus publications of all German-affiliated authors

This paper relies on a complete enumeration of the Scopus-indexed publications of researchers who have published with an affiliation address in Germany at some point during the 1996-2020 period (over eight million publications from more than one million researchers). Note that these data do not provide information about researchers who have not published in Scopus-indexed sources during the temporal window. The unit of the data is an *authorship record,* which we define as the linkage between an author affiliation and a publication.

The raw data are pre-processed to ensure that we are performing a reliable analysis of mobility events based on the changes in affiliation addresses. The pre-processing steps involve the adoption of an unsupervised machine learning algorithm for disambiguating authors and a neural network algorithm for handling missing values (Miranda- González et al., 2020; Subbotin & Aref, 2020). In our dataset, a large majority of authorship records have a country variable. However, there are 96,465 authorship records with missing country information. The missing values are systematically inferred using a neural network algorithm inspired by Miranda-González et al. (2020), which takes an affiliation address as the input and predicts the country as the output. We use a random set of one million authorship records from our dataset that have country information, and use them as training data (80%) and testing data (20%). The technical details of the development of the neural network have been explained elsewhere (Miranda-González et al. 2020). It has been demonstrated that the neural network can correctly predict the country for 98.4% of records, which is a level of accuracy we consider acceptable for predicting missing country information.

The second step of our data pre-processing helps us overcome the problems associated with using Scopus author IDs to identify unique authors. It has been shown that Scopus author IDs have high levels of *precision* and *completeness*. Precision measures the percentage of author IDs that are associated with the publications of a single individual only. Completeness measures the percentage of author IDs that are associated with all of the Scopus publications of an individual. The results of the latest evaluation of the accuracy of Scopus author IDs conducted in August 2020 showed that the precision and the completeness of Scopus author profiles are 98.3% and 90.6%, respectively (Paturi & Loktev, 2020). However, while it appears that the quality of individual-level Scopus data is sufficiently high to enable us to study the migration of researchers (Kawashima & Tomizawa, 2015; Aman, 2018), there are several notable limitations to keep in mind when using Scopus data for migration research. The precision limits in Scopus author IDs imply that 1.7% of Scopus author IDs may be associated with the publications of more than one person, which could affect the accuracy of the migration events detected through changes of affiliation countries per author ID. We overcome this problem by using an unsupervised machine learning algorithm that was inspired by a recently developed author name disambiguation method (D'Angelo and van Eck, 2020). The conservative adaptation of this method (Miranda-González et al., 2020) assumes that every two authorship records are from distinct individuals unless sufficient evidence is found to the contrary using a rule-based scoring approach and a clustering method. We first calculate the similarity score of each pair of two authorship records belonging to the same author ID. The similarity is measured based on author names, co-author names, subjects, funding information, and grant numbers. The author disambiguation algorithm makes all pairwise comparisons between authorship records with the same author ID, and creates a distance matrix based on similarities and dissimilarities in the aforementioned features for each pair of records. A clustering algorithm is then used to process the distance matrices, and to cluster similar authorship records. We then issue revised author IDs based on the resulting clusters. We use the *agglomerative clustering* algorithm from the scikit-learn Python library (Pedregosa et al., 2011) to cluster authorship records. This algorithm belongs to the family of hierarchical clustering methods. Supporting our conservative approach, it first places each record in its own cluster, and then merges pairs of clusters successively if doing so minimally increases a given linkage distance (Pedregosa et al., 2011). As well as being compatible with our conservative approach, agglomerative clustering has the advantage of offering us the flexibility to process any pairwise distance matrix.

Our author disambiguation procedures are based on the extraction of a subset of *suspicious* author IDs that are especially likely to be affected by the precision flaws of Scopus author IDs. An author ID is considered suspicious if it is associated with more than six countries or more than 292 publications (an average of more than one publication per month across a period of 24 years and four months). Based on these criteria, only 25,000 out of a total of 1.4 million author IDs are classified as suspicious. These author IDs are associated with 2,242,797 publications. After disambiguation, revised author IDs are issued for these records according to their clusters, and then they are merged with the rest of the data for the purpose of inferring migration events.

The international mobility of researchers is determined by identifying the changes in the affiliation addresses of authors across different publications over time. To detect migration events more reliably, the most frequent (mode) country(ies) of affiliation is extracted for each researcher in each year. A migration event is considered to have happened only if the mode country of affiliation changes for the researcher across different years such that the previous mode country disappears. By aggregating the movements for each pair of origin and destination countries, we can estimate the international scholarly migration flows. Accordingly, the country of academic origin is defined as the mode country during the first year of publishing. Similarly, the country of destination is defined as the mode country of the most recent year of publishing.

Based on their migration events or lack thereof, and on their academic origins and destinations, researchers are assigned to one of the following six categories (mobility types) from the perspective of the German science system:

(1) Single-paper author (having only one publication);
(2) Non-mover (having multiple publications and Germany as the only mode country);
(3) Immigrant (origin: not Germany; destination: Germany);
(4) Emigrant (origin: Germany; destination: not Germany);
(5) Return migrant (origin: Germany; destination: Germany, with international migration); and
(6) Transient (origin: not Germany; destination: not Germany, but with Germany being among the researcher's mode countries at some point).

The annual net migration rate (NMR) is calculated based on the difference between incoming and outgoing flows in a year divided by the population of research-active scholars in that year, and then expressed as a per-thousand rate. The size of the population of research-active scholars in a given year is based on the number of researchers who have listed Germany as their country of affiliation on publications within a two-year vicinity of that year. A positive NMR value can be interpreted as indicating that more researchers are entering than leaving the country under analysis.

We define the *academic age* of researchers as the number of years since their first publication (as of 2020). We calculate the average annual citation rate for each researcher by dividing their total number of citations (as of April 2020) by their academic age. This measure allows us to compare the citation performance of researchers who have different levels of experience, but are in the same field. When comparing the citations of migrant researchers from different fields, we apply a second normalization by dividing the annual citation rate by the average annual citation rate of migrant researchers in that field.

We use the *All Science Journal Classification* (ASJC) codes to incorporate the fields and disciplines (subfields) of scholarship in our analysis. The ASJC codes are based on four fields (*Heath Sciences*, *Life Sciences*, *Physical Sciences*, and *Social Sciences*). These codes are further divided into 26 academic disciplines that provide a more granular level of information about the scholarship. The field and the discipline of each researcher are determined based on the frequencies of the ASJC codes in the individual's publications. We first compute the frequencies of each field and discipline in the authorship records of an individual, and then compare them using a Z-test (Subbotin & Aref, 2020) with the mean and the standard deviation of frequencies for that field or discipline among all researchers in our dataset. The main field and discipline of each researcher are determined based on the largest Z-score that exceeds a threshold of one. If neither of the Z-scores exceeds one, we categorize the researcher as *Multidisciplinary*.

The first names of researchers can be used to infer their gender (Larivière et al., 2013; Goldstein & Stecklov, 2016b, a; Dworkin et al., 2020). This study utilizes the *Genderize* library, a simple application programming interface (API) that provides the gender most commonly associated with a given first name from a dataset of over 114 million names. After performing basic text operations (like removing middle initials from the first name field), we obtained the gender for 1,046,873 author profiles in our dataset. For the remaining profiles, we manually searched for public author information to determine the gender by checking the individual's personal homepages, curricula vitae, online profiles, biographies in publications, and other online sources. Accordingly, the genders for 6,767 additional author profiles were determined manually. Finally, the gender of 1,053,640 author profiles (75.96%) were reliably determined by either algorithmic or manual gender detection. For our analyses that involve gender (e.g., measuring gender ratios), we set aside the 24.04% of author profiles whose gender cannot be determined either algorithmically or manually. The binary genders inferred and used in our analysis do not refer directly to the sex of the researchers, assigned at birth or self-chosen; nor do they refer to the socially assigned or self-chosen genders of the authors.

Results

We look at citation-based performance by discipline across different mobility types. We also provide descriptive results on common countries of origin and destination, and calculate the annual net migration rates of researchers for Germany. Furthermore, the discipline-normalized citation performance of the researchers is evaluated and compared for the common origin and destination countries. Finally, we provide detailed results on the gender ratios of researchers by discipline, and compare the gender ratios among mobile researchers and all researchers.

Mobility types and annual citation rates by discipline

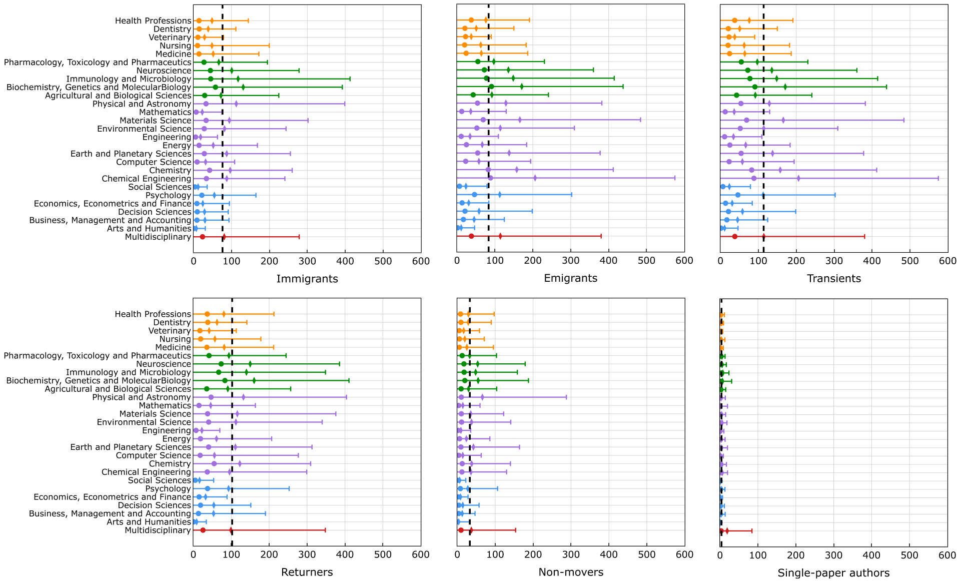
Among the researchers who have Germany as a mode country for at least one year, 44.01% are single-paper authors, 42.08% are non-movers, and only 13.91% are internationally mobile. More details are provided in Table 1.

Table 1. Number and proportion of researchers by mobility type.

	Single-paper author	*Non-mover*	*Immigrant*	*Emigrant*	*Return migrant*	*Transient*
Count	610,449	583,679	51,135	63,003	37,078	41,722
%	44.01%	42.08%	3.69%	4.54%	2.67%	3.01%

Figure 1 shows the average, median, and standard deviations of annual citation rates by discipline for researchers of each mobility type. There are substantial disparities in the annual citation rates of these researchers, as the large standard deviations indicate. As expected, we find that there are notable differences in the annual citation rates of researchers across different disciplines (Marmolejo-Leyva et al., 2015; Bedenlier, 2017; Horta et al., 2019). On average, researchers in the social sciences (colored in blue) receive fewer citations and researchers in the life sciences (green) have more citations than their counterparts in other fields.

When we look at differences in citation rates by mobility type, we can see that mobile researchers markedly outperform non-movers, even after separating single-paper authors. Compared to non-movers, mobile researchers have much higher annual citation rates for almost all disciplines. Thus, even though they represent a minority of the researchers in Germany, mobile researchers make substantial scientific contributions in terms of receiving citations. Among the internationally mobile researchers, transients and return migrants have higher annual citation rates than immigrants and emigrants. This pattern may be explained in part by the more complex mobility trajectories of transients and return migrants.



Immigrants

Emigrants

Transients

Returners

Non-movers

Single-paper authors

Figure 1. Average (diamond marker), median (circle marker), and standard deviation of annual citations by discipline and mobility type. The vertical dashed lines show the mean across all disciplines. Magnify all figures on the screen for higher resolution and more details.

Flows, origins, and destinations

Figure 2 illustrates the total migration flows of researchers from and to Germany during the 1996-2020 period. By far, the largest academic migration inflows and outflows are with the US, as various studies have documented (Maier et al., 2007; Weinberg, 2009; Hunter et al., 2009; Franzoni et al., 2012; Guthrie et al., 2017). For Germany, the US is also the most common resource hub of highly skilled individuals, due to the large number of researchers in the US and their high mobility potential (Schiller & Cordes, 2016). After the US, the common academic origins and destinations for Germany-affiliated migrant researchers are the UK in a distant second, followed by Switzerland. Most of the other common origin and destination countries are in Europe, which indicates that geographical proximity is also relevant. Moreover, EU policies might have played a role in facilitating large migration flows of scholars within Europe. For example, the *Bologna Process,* which was introduced in the 2000s, was designed to ensure comparability in the standards and the quality of higher education qualifications in Europe (Teichler, 2015). A more recent example is the *IPID4all* program, which was launched in 2014 with the aim of creating an attractive environment in Germany for young international researchers (IDEA Consult et al., 2017). Such polices are helping to drive and facilitate academic mobility by providing researchers with more opportunities for academic exchange and communication. In addition, there are a number of countries that are exporting substantial numbers of published researchers to Germany, but that are not equally attractive destinations for researchers from Germany, including Russia and India.

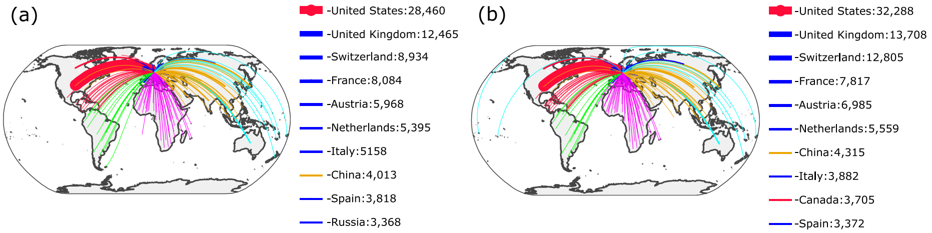


Figure 2. Migration flows of researchers to Germany (a) and from Germany (b) over the 1996-2020 period.

Net migration rates

Figure 3 illustrates the shifting annual net migration rates of researchers in Germany over the 1998-2017 period (other years are not reported due to boundary effects). The data suggest that for nearly the entire two decades of our analysis, the outgoing flow of published researchers exceeded the incoming flow. The lowest NMR value is for 2008, at -10.52 per 1,000 researchers. Note that the NMR in the general population also reached its lowest value in 2008, which was during the global economic crisis. However, in contrast to migration among researchers, the net migration rate for the general population as a whole was consistently positive throughout these years. After 2008, the NMR for researchers displayed a generally increasing trend until 2014, when it reached a peak of +1.56. In the most recent years, the NMR for researchers followed a generally decreasing trend. Our finding of a generally negative NMR for scholars is consistent with the results of studies by Schiller et al., which compared the birthplaces and workplaces of researchers, and reported that Germany has suffered a net loss of 28% of researchers due to international mobility (Schiller & Revilla Diez, 2008; Schiller & Cordes, 2016). The recent trends in migration among scholars contrasts with the NMR in the general population, which indicates that since 2008, Germany has increasingly become an immigrant destination. Indeed, since 2012, Germany has admitted large numbers of asylum-seekers.

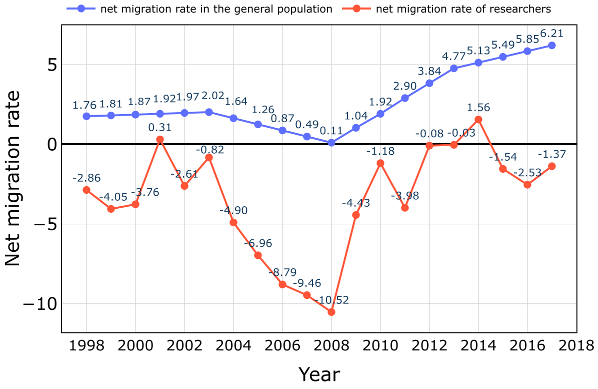


Figure 3. Net migration rates for researchers (in red) and for the general population (in blue).

Citation-based performance by geography

We previously examined the disparities in the citation-based performance of researchers of different mobility types (Fig. 1). In this subsection, we analyze the citation performance of immigrant and emigrant researchers by country. First, we divide the annual citation rate of each migrant researcher by the average of all migrant researchers in the respective discipline to obtain a discipline-normalized annual citation measure. Based on the three quantiles of the resulting distribution for all migrant researchers, we have distinguished three equally sized citation groups: low, medium, and high. Accordingly, the discipline-normalized annual citation rate is between 0.18 and 0.73 for migrant researchers in the medium group, it is below 0.18 for those in the low group, and it is above 0.73 for those in the high group.

Figure 4 shows the composition of the citation groups among the emigrants and immigrants for 30 countries.

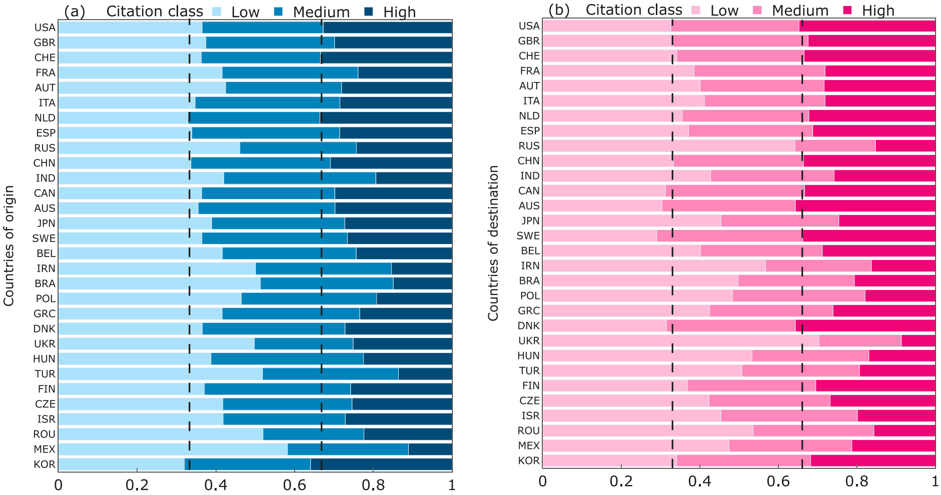


Figure 4. Citation class composition of immigrants by origin (a) and emigrants by destination (b). Vertical gridlines are shown to enable a comparison to the average migrant researcher, who is equally likely to belong to any of the three citation classes.

Among emigrants, scholars moving to Denmark, Austria, and Sweden are more likely than other emigrants to belong to the high citation group. Among the common destinations for emigrants, Sweden, Austria, and Denmark rank ninth, 10th, and 13th, respectively. Among immigrants, those who come from South Korea and the Netherlands are more likely than those who come from other countries to belong to the high citation group.

When we look at the numbers of immigrant and emigrants from the 10 countries with the largest flows, we observe that for most citation groups and country combinations, there are more emigrants than immigrants. This finding is also consistent with the overall negative net migration rate that we observed in Figure 3. For the US, the UK, and Switzerland (which have the three largest flows), the number of emigrants is considerably higher than the number of immigrants. Conversely, while Russia, Italy, and Spain, have comparable migration flows with Germany, they send more published researchers to Germany than they receive from Germany.

Gender disparities among migrant researchers by discipline

Figure 5 shows on the X-axis that almost all disciplines in Germany are dominated by male researchers. Especially for the disciplines in the physical sciences, like engineering, physics, and astronomy, the male-to-female ratio exceeds eight. The only disciplines for which the gender ratio is close to unity are veterinary sciences, psychology, and immunology and microbiology.

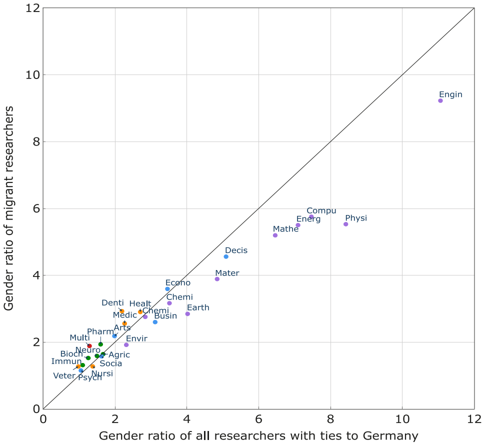


Figure 5. Male-to-female gender ratios of migrants and all researchers by discipline

Given these observations, we examine the gender ratios among the migrant researchers of each discipline, and then compare them to the corresponding ratios among all German-affiliated researchers in Figure 5. The most striking finding is that in disciplines that have a gender ratio above four among all researchers, the gender disparity is less severe among the migrant researchers (see the data points under the 45° line, and especially all of the disciplines with ratios above four). In contrast, for most disciplines in the life sciences (green dots), the male-to-female ratio is more balanced than it is in the physical sciences, and it is slightly higher among migrant researchers.

Discussion and future directions

As researchers play a unique role in innovation and economic growth, and are among the most mobile members of the occupational structure, most countries impose few, if any, limits on their mobility. Thus, it is crucial that we understand whether and, if so, how countries like Germany are situated in the global migration flows of published researchers (Docquier & Marfouk, 2004; Schiller & Cordes, 2016; Aref et al., 2019). We used Scopus bibliometric data (Falagas et al., 2008; Mongeon & Paul-Hus, 2016) to extrapolate the key variables and to infer the geographic locations of researchers. In addition, we drew from these data information on individual researchers, including on their scientific discipline, gender, and citation performance. Based on this information, we created five distinct mobility categories for researchers who had multiple publications and were exposed to mobility. These categories accounted for 55.99% of the 1.4 million author profiles identified in the Scopus data.

Algorithmically pre-processed and author-disambiguated Scopus data enabled us to provide uniquely rich, albeit preliminary, insights into Germany's role in the global system of migration among published researchers. Examining the 24-year period from 1996 to 2020, we demonstrated that internationally mobile researchers accounted for nearly 14% of the population of published researchers with ties to Germany. This group of internationally mobile researchers made disproportionately large contributions, as their differences in citation-based performance compared very favorably to those of non-movers in Germany.

The sheer size of our census of German-affiliated researcher data allowed us to differentiate the population of internationally mobile researchers according to their countries of origin and destination. Thus, these data provided us with a unique and detailed mapping of the key countries associated with academic mobility to and from Germany. We found that the US, the UK, and Switzerland were the three largest origin countries of published researchers immigrating to Germany. Our findings also showed that Germany has gone beyond reciprocating these levels of migration, with higher numbers of researchers emigrating from Germany to these countries. While these three countries were the primary destinations and accounted for 58% of all scholarly migration in this period, they were not necessarily the most significant in terms of their citation performance. Researchers who emigrated to Denmark, Sweden, and Austria and researchers who immigrated from South Korea and the Netherlands were more likely to be highly cited than other German-affiliated immigrants and emigrants.

Our data also allowed us to explore levels of gender disparities in the German science system, to investigate how these inequalities varied between migrant and *stationary researchers* (non-movers and single-paper authors), and to explore the heterogeneity in levels of gender inequality across scientific fields. We observed that the gender ratios (males to females) were nearly equal in disciplines such as veterinary sciences and psychology, whereas in fields such as engineering and physics and astronomy, there were six to eight times more men than women. While some of these patterns have been well-documented in the literature (Larivière et al., 2013; Macaluso et al., 2016; Sugimoto et al., 2015; IDEA Consult et al., 2017), our results highlight the key role that international mobility may play in helping to moderate some of the most extreme gender disparities. We found that female researchers were better represented among migrants than among stationary researchers in the most gender-imbalanced disciplines. This observation allows us to speculate that migration could further counterbalance the extremely gendered nature of certain disciplines in Germany. The results also highlight that more can be done to support female scholars wishing to remain in Germany, and to encourage the immigration of female scholars in fields that are heavily unequal, such as engineering, physics and astronomy, computer science, energy, and mathematics.

The analysis enabled us to uncover – within the quality limitations of the bibliometric data and our pre-processing approach for handling missing values and disambiguating authors – new aspects of the academic life course, and to show how it is connected to scholarly migration. Ultimately, our analysis helps to shed light on Germany as both a hub and a through station on the international map of academic mobility. The preliminary questions explored here offer some initial insights for policymakers, as they improve our understanding of the role of migrant researchers in the German science system, and their demographic composition. Our analysis contributes to the science of science by providing relevant evaluation measurements of the mobility of researchers from all disciplines based on their geography, gender, and citation performance.

Our study opens a number of new directions for further research. Building on this work, we intend to move from using an aggregate perspective to applying a micro-level perspective in order to gain further insights into the relationship between migration and scientific performance. This approach will be particularly useful for studying how women are affected, especially in fields where they are severely under-represented, and for examining the impact of specific national policies in Germany that aim to improve the gender balance in academic fields. Ultimately, by understanding the role of migration both into and out of Germany in each field and discipline, we can provide additional insights into researchers' characteristics, mobility behavior, and performance. Thus, our study represents a key step toward gaining a better understanding of migration among scholars.

Acknowledgments

The authors highly appreciate the technical support from Tom Theile and the comments from Aliakbar Akbaritabar. This study has been funded by the German Academic Exchange Service with funds from the Federal Ministry of Education and Research. This study has received access to the bibliometric data through the project "Kompetenzzentrum Bibliometrie," and the authors acknowledge their funder BMBF (funding identification number 01PQ17001).

References

Aman, V. (2016). How collaboration impacts citation flows within the German science system.

Scientometrics, 109(3), 2195-2216.

Aman, V. (2018). Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. *Scientometrics*, 117(2), 705-720.

Aref, S., Zagheni, E., & West, J. (2019). The demography of the peripatetic researcher: Evidence on highly mobile scholars from the Web of Science. In *International Conference on Social Informatics* (pp. 55-65). Springer.

Bauder, H. (2015). The international mobility of academics: A labour market perspective. *International Migration*, 53(1), 83-96.

Bedenlier, S. (2017). Internationalization within higher education and its influence on faculty: experiences of Turkish academic staff. *Journal of Research in International Education*, 16(2), 185-196.

Børing, P., Flanagan, K., Gagliardi, D., Kaloudis, A., & Karakasidou, A. (2015). International mobility: Findings from a survey of researchers in the EU. *Science and Public Policy*,42(6), 811-826.

European Commission. (2012). A reinforced European research area partnership for excellence and growth. *Report for COM (2012).* Brussels: European Commission.

Docquier, F. & Marfouk, A. (2004). *Measuring the International Mobility of Skilled Workers (1990-2000): Release1.0.* The World Bank.

Dubois, P., Rochet, J. C., & Schlenker, J. M. (2014). Productivity and mobility in academic research: Evidence from mathematicians. *Scientometrics*, 98(3), 1669-1701.

Dworkin, J. D., Linn, K. A., Teich, E. G., Zurn, P., Shinohara, R. T., & Bassett, D. S. (2020). The extent and drivers of gender imbalance in neuroscience reference lists. *Nature neuroscience*, *23*(8), 918-926.

D'Angelo, C. A. & van Eck, N. J. (2020). Collecting large-scale publication data at the level of individual researchers: A practical proposal for author name disambiguation. *Scientometrics*, 123, 883-907.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal*, 22(2), 338-342.

Franzoni, C., Giuseppe, S., & Stephan, P. (2012). Patterns of international mobility of researchers: Evidence from the GlobSci survey. In *International Schumpeter Society Conference*, 14, 2-5.

Franzoni, C., Scellato, G., & Stephan, P. (2014). The mover's advantage: The superior performance of migrant scientists. *Economics Letters*, 122(1), 89-93.

Franzoni, C., Scellato, G., & Stephan, P. (2015). Chapter 2 - International mobility of research scientists: Lessons from GlobSci. In Geuna, A. (Eds), *Global Mobility of Research Scientists* (pp. 35-65). Academic Press.

Gibson, J. & McKenzie, D. (2014). Scientific mobility and knowledge networks in high emigration countries: Evidence from the pacific. *Research policy*, 43(9), 1486-1495.

Giousmpasoglou, C. & Koniordos, S. (2017). Brain drain in higher education in Europe: Current trends and future perspectives. In G. Charalampos, V. Paliktzoglou & E. Marinakou (Eds), *Brain Drain in HigherEducation: TheCaseoftheSouthernEuropeanCountries andIreland* (pp. 229-262). New York: Nova Science Publishers.

Goldstein, J. R. & Stecklov, G. (2016a). From Patrick to John F.: Ethnic names and occupational success in the last era of mass migration. *American Sociological Review*, 81(1), 85-106.

Goldstein, J. R. & Stecklov, G. (2016b). Measuring assimilation of the first waveof Mexicanimmigrants to theUnited States, 1910–1940. In*2016Annual Meeting*. PAA.

Guthrie, S., Lichten, C. A., Harte, E., Parks, S., & Wooding, S. (2017). *International mobility of researchers: A survey of researchers in the UK*. RAND Corporation.

Horta, H., Jung, J., & Santos, J. M. (2019). Mobility and research performance ofacademicsincity-based highereducationsystems.*HigherEducationPolicy*, pp.1-22.

Hunter, R. S., Oswald, A. J., & Charlton, B. G. (2009). The elite brain drain. *The Economic Journal*, 119(538), F231-F251.

IDEA Consult (2013). *Support for continued data collection and analysis concerning mobility patterns and career paths of researchers*. *Final report MORE2*. Brussels.

IDEA Consult., WIFO., & Technopolis. (2017). *MORE3 study: Support data collection and analysis concerning mobility patterns and career paths of researchers*. Brussels.

Ioannidis, J. P. A. (2004). Global estimates of high-level brain drain and deficit. *The FASEB Journal,* 18(9), 936-939.

Jöns, H. (2009). 'Brain circulation' and transnational knowledge networks: Studying long-term effects of academic mobility to Germany, 1954-2000. *Global Networks*, 9(3), 315-338.

Kawashima, H. & Tomizawa, H. (2015). Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics*, 103(3), 1061-1071.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, 504(7479), 211.

Lörz, M., Netz, N., & Quast, H. (2016). Why do students from underprivileged families less often intend to study abroad? *Higher Education*, 72(2), 153-174.

Macaluso, B., Larivière, V., Sugimoto, T., & Sugimoto, C. R. (2016). Is science built on the shoulders of women? A study of gender differences in contributorship. *Academic Medicine*, 91(8), 1136-1142.

Maier, G., Kurka, B., & Trippl, M. (2007). Knowledge spillover agents and regional development: Spatial distribution and mobility of star scientists. *DYNREG (Dynamic Regions in a Knowledge-Driven Global Economy)*, 17, 35.

Marmolejo-Leyva, R., Perez-Angon, M. A., & Russell, J. M. (2015). Mobility and international collaboration: case of the Mexican scientific diaspora. *PloS one*, 10(6), e0126720.

Miranda-González, A., Aref, S., Theile, T., & Zagheni, E. (2020). Scholarly migration within Mexico: Analyzing internal migration among researchers using Scopus longitudinal bibliometric data. *EPJ Data Science.* 9(1), 1-26

Moed, H. F. & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics*, 101(3), 1987-2001.

Mongeon, P. & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, 106(1), 213-228.

Netz, N. & Finger, C. (2016). New horizontal inequalities in German higher education? social selectivity of studying abroad between 1991 and 2012. *Sociology of Education*, 89(2), 79-98.

Netz, N. & Grüttner, M. (2020). Does the effect of studying abroad on labour income vary by graduates' social origin? evidence from Germany. *Higher Education*, pp. 1-23.

Netz, N., Hampel, S., & Aman, V. (2020). What effects does international mobility have on scientists' careers? a systematic review. *Research Evaluation*, 29(3), 327-351.

Netz, N. & Jaksztat, S. (2017). Explaining scientists' plans for international mobility from a life course perspective. *Research in Higher Education*, 58(5), 497-519.

Parey, M., Ruhose, J., Waldinger, F., & Netz, N. (2017). The selection of high-skilled emigrants. *Review of Economics and Statistics*, 99(5), 776-792.

Paturi, M. & Loktev, A. (2020). The best gets better: Scopus data quality measured. In *Pure International Conference*. https://brighttalk.com/webcast/13819/456949/prcn2020-the-best-gets-better-scopus-data-quality-measured

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825-2830.

Rosenfeld, R. A. & Jones, J. A. (1987). Patterns and effects of geographic mobility for academic women and men. *The Journal of Higher Education*, 58(5), 493-515.

Schiller, D. & Cordes, A. (2016). Measuring researcher mobility. In *OECD Blue Sky Forum*, pp. 24.

Schiller, D. & Revilla Diez, J. (2008). Mobile star scientists as regional knowledge spillover agents. In *IAREG Working Paper*, 11,3-33.

Subbotin, A. & Aref, S. (2020). Brain drain and brain gain in Russia: Analyzing international mobility of researchers by discipline using Scopus bibliometric data 1996-2020. *ArXiv.* https://arxiv.org/pdf/2008.03129.

Sugimoto, C. R., Ni, C., & Larivière, V. (2015). On the relationship between gender disparities in scholarly communication and country-level development indicators. *Science and Public Policy*, 42(6), 789-810.

Teichler, U. (2015). Academic mobility and migration: What we know and what we do not know. *European Review*, 23, S6-S37.

Weinberg, B. A. (2009). An assessment of British science over the twentieth century. *The Economic Journal*, 119(538), F252-F269.

The evolution of interdisciplinarity in five social sciences and humanities disciplines: relations to impact and disruptiveness

Hongyu Zhou[1], Raf Guns[1] and Tim C. E. Engels[1]

[1] *Hongyu.Zhou@uantwerpen.be, Raf.Guns@uantwerpen.be, Tim.Engels@uantwerpen.be*

Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, University of Antwerp Middelheimlaan 1, 2020 Antwerp (Belgium)

Abstract

It is generally believed that the tide of interdisciplinarity is rising and becomes increasingly prevalent among various disciplines in natural and biomedical sciences. However, for the social sciences and humanities (SSH) limited evidence supports such a statement from bibliometric perspectives. Also, it has seldom been quantified how interdisciplinarity and its various aspects evolve over time. This paper analyzes the evolution of interdisciplinarity focusing on two aspects, namely process and outcomes, to draw a comprehensive trajectory of interdisciplinarity for five SSH disciplines over 50 years. We find that research in each of these five SSH disciplines is broadening its knowledge base by involving more disciplines yet is at the same time shifting towards further specialization. Interdisciplinarity is found to be positively correlated with citation impact and visibility and becomes stronger as the citation window widens up. The disruptiveness of publications, however, is negatively correlated with the level of interdisciplinarity and increasingly so over time.

Introduction

Interdisciplinary research (IDR) emerges to tackle the complex and societal-pressing problems that cannot be truly resolved by a single discipline (Carayol & Thi, 2005; Frodeman & Mitcham, 2016). To closely monitor and evaluate the supporting initiatives and understanding the status and mechanisms behind IDR, several recent studies are devoted to examining its different aspects (Rousseau et al., 2019), such as input (e.g. disciplinary diversity in team assembly; Schummer, 2004; Zhang et al., 2018), process (e.g. disciplinary diversity in references; Mugabushaka et al., 2016; Porter et al., 2007), outputs (e.g. topic diversity in abstracts; Bu et al., 2020), and outcomes (e.g. research impact; Larivière et al., 2015; Larivière & Gingras, 2010; Szell et al., 2018; J. Wang et al., 2015). Among all, quantitative measurements and indicators to evaluate the intensity of IDR processes, i.e., how interdisciplinary one's research is, is one of the most discussed focal research topics (Wagner et al., 2011). Although criticized as confusing and unsatisfying to achieve universally convergent assessment for precise decision-making (Q. Wang & Schneider, 2019), these indicators can still assist us to get a glimpse of IDR as a social phenomenon and to interpret it from a macro perspective.

One of the frequently referenced macroscopic statements by researchers, policy-makers, and the media is that "science is becoming more interdisciplinary". A consensus seems to have formed that scientists inhabiting dissimilar knowledge bases or mastering different skills have been crossing disciplinary borders and collaborating more frequently and with unconventional partners; this leads to more interdisciplinary and scientifically significant outcomes. Empirical evidence, however, is still limited to disciplines from STEM and biomedical sciences. Temporal change in IDR in the social sciences and humanities has not been studied so far.

In this study, we analyse the evolution of interdisciplinarity in five SSH disciplines over half a century and examine the possible impact such change in IDR, if any, might produce on the outcome of scientific research. The rest of the paper is organized as follows: we first introduce the dataset and methodology we adopt. The next section presents the results and discussion and the last section concludes.

Data

The Microsoft Academic Graph (MAG) dataset is adopted in our empirical study; previous studies have shown that MAG is a viable source for scholarly communication research in terms of coverage (Paszcza, 2016) and the completeness of citation and metadata (Thelwall, 2017). As Microsoft continues to improve the coverage, design, and accessibility of MAG, it has become one of the most promising bibliographic datasets and is more frequently employed by those who study research dynamics quantitatively (Kong et al., 2020; Ma et al., 2020).

MAG distinguish from other bibliographic databases in that it adopts a bottom-up approach for the field categorization process (K. Wang et al., 2020), as opposed to, for instance, Web of Science, which uses existing journal categories to classify publications (called a top-down or classification-based approach, see Wagner et al., 2011). Different from previous bottom-up classification methods that tend to adopt co-citations, co-words, and/or bibliographic couplings (Wagner et al., 2011), MAG quantifies the semantic distance between two textual paragraphs representing two certain publications and then clusters the retrieved semantic representations to form the basis of concepts, which are de facto fields, domains, or disciplines in practice. Six levels of concepts are clustered automatically on different granularities. The top two levels of concepts (L0 and L1) are manually defined into a unique hierarchical structure to be consistent with most of the categorization systems (K. Wang et al., 2020), where L0 is comprised of 19 fields (e.g., physics, chemistry, and economics) and L1 consists of 294 subfields (e.g., theoretical physics, biochemistry, and macroeconomics).

The disciplines we use as case studies in this paper, i.e. Anthropology, Applied Psychology, Linguistics, Library Science, and Macroeconomics, are situated in the L1 level of MAG's category setting. Therefore, L1 is also used in the categorization of individual publications and their corresponding references to achieve consistency.

We recognized 300,559 journal or conference publications between 1960 and 2009 under the field category Anthropology (62,619), Applied Psychology (116,868), Linguistics (65,557), Library Science (17,678) and Macroeconomics (39,481). Publications labeled for more than one category were assigned to each category.

Method

We investigate the evolution of IDR focusing on two aspects, namely IDR process, and IDR outcome.

IDR Process

Stirling (2007) pointed out that diversity consists of three basic concepts, namely variety, balance, and disparity, each of which is a necessary but insufficient property of diversity as a whole. This notion and their generic indicator of diversity were then introduced and modified by Rafols and Meyer (2010) to Information Science as a quantitative measurement of knowledge integration to infer interdisciplinarity. A significant proportion of research is devoted to devising indicators that integrate two or three factors (dimensions) of diversity to achieve a reliable metric and assess or compare interdisciplinarity for different entities. In this study, we try to work as comprehensively and detailed as possible so that information loss caused by dimension reduction or integration can be minimized. Therefore, to quantify the intensity and evolution of IDR processes, we employed both single-factor (variety, balance, and disparity themselves; Stirling, 2007) and multi-factor measurements: Rao-Stirling (RS) diversity (Rafols & Meyer, 2010; Stirling, 2007), DIV (Leydesdorff et al., 2019), and $^2D^s$ (Zhang, Rousseau, & Glänzel, 2016) that involve two or all three single factors. We believe that the involvement of single-factor measurements may provide more implications that directly point to practical aspects of IDR, for instance, the number of disciplines referenced, and that the multi-factor measurements shed insight into the evolution of IDR from a more comprehensive perspective.

Table 1 provides notations and mathematical definitions of each indicator we employed in this study. Variety (n_c) is operationalized as the number of disciplines referenced for each publication, which reveals information regarding the **broadness of the knowledge base** in this study. Its variant, relative variety (n_c/N), is used in DIV representing variety in a relative scale; this variant is defined as variety divided by the number of categories in total. Balance (B), representing the **evenness of the knowledge base**, is set to be $1 - Gini_c$ where B = 1 indicates maximum evenness and B = 0 shows extreme imbalance. Here, $Gini_c$ represents the Gini coefficient of the distribution of disciplines in references. Disparity captures the average dissimilarity (or distance, explained further on) between every two disciplines referenced for each publication, which can be utilized to examine the cognitive distance and **heterogeneity of the knowledge base**.

Table 1. Selected measures of IDR.

Notation			
n_c	number of disciplines referenced	d_{ij}	dissimilarity between categories *i* and *j*
p_i	proportion of elements in category *i*	N	number of categories in total
x_i	number of references to the i-the category in an ascending order		
Indices			
Variety	n_c	Rao-Stirling (RS)	$\sum_{i,j} d_{ij}(p_i p_j)$
Balance (B)	$1 - Gini_c = \quad 1 - \frac{\sum(2i - n_c - 1)x_i}{\sum_{n_c} x_i}$	DIV	$(n_c/N) * B * D$
Disparity (D)	$\frac{\sum_{i \neq j} d_{ij}}{[n_c * (n_c - 1)]}$	$^2D^s$	$1/(1 - RS)$

Three indicators integrating a part of or all the above-mentioned three factors are also employed (i.e., multi-factor indicators as aforementioned). DIV is the multiplication of relative variety, balance, and disparity ranging from zero to one. In terms of RS, this indicator calculates the sum of distances between every two disciplines referenced multiplied by the product of proportions each discipline accounts for in reference. $^2D^s$can be regarded as a variant of RS that possesses greater discriminatory power, satisfying the properties proposed in Leinster and Cobbold (2012). Like RS, this indicator employs similarity among categories instead of disparity directly.

Four out of six measures (i.e. disparity, RS, DIV, and ${}^{2}D^{s}$) employ the dissimilarity between two categories d_{ij} in their calculation, which is operationalized as (1 - similarity) in practice, as shown to be valid and efficient in Zhang, Rousseau, and Glänzel (2016). The temporal perspective of this paper makes a few modifications to the cosine similarity necessary, that is, the application of a time window on similarity calculation. As the distance or reference strength among disciplines may be changing over time (Frank et al., 2019), potential structural changes to the similarity matrix itself cannot be ignored when performing temporal analysis. To account for this, we construct ten similarity matrices with a five-year time window each. This yields the following equation:

$$d_{ij} = 1 - \frac{R_{ij}+R_{ji}}{\sqrt{\left(TC_i^t + TR_i\right)\left(TC_j^t + TR_j\right)}} \tag{1}$$

where i and j refer to two sets of publications from two different categories published during the period t, R_{ij} denotes the number of times set i publications cite set j publications, TC_i^t denotes the total number of citations set i publications received during the period t, and TR_i denotes the total number of references initiated by papers from set i.

The handling of multi-labeling in the categorization of publications is also a tricky issue when calculating balance, disparity, and RS as counting frequencies of categories for references is required. Upon our examination, more than 50% of the publications in our dataset are labeled with more than one category. To address this issue, we used fractional counting to quantify the relative frequencies. For each multi-labeled reference, we set the frequency of each category labeled to $1/m$ where m equals the number of unique categories associated with this reference and then sum all the category frequencies for each reference to form the category frequency for the publication. For example, a publication with two references A and B, where A is labeled with both "Linguistics" and "Literature", and B is labeled with "Linguistics" and "Natural Language Processing (NLP)", will have an overall category frequency distribution as follows: Linguistics 1; Literature 0.5; and NLP 0.5.

IDR outcome

Furthermore, the evolution of IDR outcome is investigated focusing on two aspects, namely **citation impact** and **disruptiveness**. 2-year, 5-year, and 10-year citations are calculated to reflect the academic significance and visibility over time. Disruptiveness, proposed in Funk & Owen-Smith (2017) and applied in Wu, Wang, and Evans (2019), captures the level of disruption a publication contributes by calculating the percentage of the net increase of citing papers it invites to an existing local citation network. As shown in equation (2), for a certain paper P, n_i denotes the number of its forward citations received during time t that did not cite any of P's backward citations, n_j represents publications published during time t citing both P and its backward citations, and n_k captures publications that only cited P's backward citations instead of P itself.

$$D_t = \left(n_i - n_j\right)/\left(n_i + n_j + n_k\right) \tag{2}$$

Results and discussion

The evolution of IDR process

Broadness of knowledge base

Figure 2a illustrates the evolution of variety for five disciplines in 50 years (1960-2009). Each subplot represents a discipline and each solid curve shows the probability distribution of variety for all publications in that discipline in the corresponding period. The vertical dashed lines denote the mean of variety for each period. The same color for a solid curve and a dashed line means that they are describing the same period.

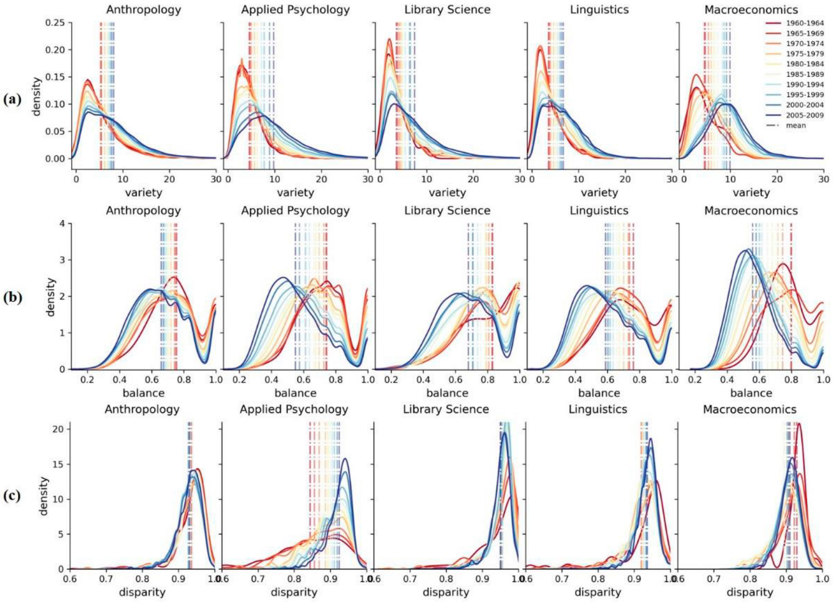


Figure 2. Evolution of variety, balance, and disparity in Anthropology, Applied Psychology, Library Science, Linguistics and Macroeconomics.

The dashed curves show sustained growth in terms of the central tendency, i.e. mean values, of variety for all five disciplines, but also illustrate a few discipline-wise differences. Applied Psychology and Macroeconomics, for instance, possess the biggest rise (from 5 to 10), which makes them the most interdisciplinary fields in terms of the broadness of the knowledge base. A rather moderate increase is associated with Linguistics and Anthropology. On the other hand, the growth of mean variety in Applied Psychology and Library Science seems to be accelerating as the gap between adjacent mean values (i.e., the distance between two adjacent vertical, dashed lines) keeps widening up over time. Such a phenomenon is missing or nebulous in the other disciplines.

The dominance of low variety publications has weakened over time as opposed to the rise of high variety publications. The rise of right tails is significant in all five disciplines which indicates an increasing percentage of high variety publications in each year. On the other hand, the right-shifted peak (mode) can also be spotted in Applied Psychology and Macroeconomics and illustrates that an obvious shift towards a diverse knowledge base for research in recent years.

The aforementioned findings capture the ever-increasing broadness of the knowledge base for research in SSH, more significantly in application-oriented disciplines. Furthermore, such multi-discipline-sourced studies have become the main force of science production. Clearly, researchers in SSH are more eager and proactive to absorb external knowledge or skills to advance their own study.

Evenness of knowledge base

In this paper, balance (B) is operationalized as $1 - Gini_c$ where B = 1 indicates the maximum evenness (e.g., an array [2,2,2,2] that has the same value for all elements has a value of balance equal to one) and B = 0 indicates the maximum imbalance. If the allocation of references to each category is evenly distributed, the authors have absorbed knowledge equally from various disciplines and treat each discipline as an equally significant part of the knowledge base to their study, hence indicating diversity and interdisciplinarity. On the contrary, if references are primarily devoted to one or a few disciplines, we can assume that certain specializations or sets of specializations take place in the study, yielding a low level of balance and less interdisciplinarity. A special and contradictory case occurs for papers that only have one category (discipline) in their references. This type of publications show a high specialization and low balance; yet, based on the formula of calculating B, its balance equals one which indicates maximum diversity. To handle this issue, in this paper, we exclude all papers with variety equal to one when examining balance. As such, 9.09% of publications for Linguistics, 4.08% for Applied Psychology, 6.84% for Anthropology, 9.62% for Library Science, and 1.57% for Macroeconomics are excluded.

As shown in Figure 2b, the distribution of balance exhibits multiple peaks and skewness in distribution. Publications with B =1 account for the right-most peak in all curves. The other peaks are associated with publications having a high balance in the earlier period and a low balance in recent years. It is also worth noticing that the skewness of the distribution also changes over time: In Linguistics, Applied Psychology, and Macroeconomics, the distribution is left-skewed in earlier years and shifted to right-skewed in recent years.

Contrary to variety, the mean of balance, as shown in vertical dashed lines, exhibits a decreasing trend for all disciplines. Macroeconomics achieved the largest drop of around 30.1% for 50 years, followed by Applied Psychology with a 26.0% decrease. The smallest drop can be found in Anthropology whose values for mean balance are reduced only by 12.6% during the same period. If we recall the results from the last section, we can see that disciplines which decreased the most in balance also achieved the biggest increase in terms of variety. One interpretation is that disciplines that tend to acquire knowledge from more peer disciplines do not always treat them as equally significant or relevant partners. Even though more disciplines are invited to their knowledge base, researchers from Applied Psychology and Macroeconomics are still heavily and unevenly reliant on a few disciplines.

The decreasing trend of balance is tightly related to the specialization of research (Foster et al., 2015). The formation of sub-disciplines or research topics might result in clusters of disciplines that are always regarded as the most significant knowledge base and more frequently and intensely referenced together thus yielding a dominant knowledge combination. Such tendency indicates that SSH researchers tend to have a clearer and more strategic agenda in terms of how to situate their research and how to learn from their peer scientists. Furthermore, we suppose that the increasing richness in knowledge base (i.e., variety) might as well be associated with the decrease in evenness (i.e., balance) to some extent. The occurrences of new disciplines (an increase of the value of variety) in the knowledge base might be naturally weak in intensity and proportions, which leads to an imbalanced knowledge base.

Heterogeneity of knowledge base

Disparity is operationalized as the mean distance among each pair of disciplines referenced. If only one discipline is referenced in a certain publication, its value of disparity is undefined as no "distance" can be defined or calculated. Therefore, we intentionally exclude this set of publications since there is no added information other than the decrease of the share of mono-source publications, which was already discussed in previous sub-sections

Figure 2c shows the evolution of disparity for our three selected disciplines, where one can see that the range of disparity looks similar, from ~0.60 to 1.00. Yet, the change of mean disparity varies fundamentally among all five. We observe an increase in Applied Psychology and Linguistics (9.6%, and 1.5%, respectively). Macroeconomics, on the other hand, has experienced a minor decrease in terms of mean disparity over the selected period. The other two disciplines, namely Anthropology and Library Science, fluctuate and remain at a similar level of disparity throughout the 50-year time window.

Besides changes in the mean values, we would also like to discuss the change in overall distributions of disparity, for which Applied Psychology exhibited more dramatic changes than the others. In the case of Applied Psychology, two clear and significant changes can be spotted, namely the formation of high peaks and the weakening in the left tail. Both indicate that researchers in Applied Psychology are referencing more "remote" or previously less connected disciplines to constitute their knowledge bases.

The decrease regarding the proportion of low variety publications illustrates a general tendency in Applied Psychology that researchers from this domain are aiming to voluntarily absorb and borrow knowledge or skills from more "remote" disciplines and the cognitive distance within their knowledge base continues to widen.

Integrated IDR trends

We calculated the evolution of integrated IDR using three indicators, namely DIV, ${}^{2}D^{s}$, and Rao-Stirling (RS) to investigate how the diversity of references as a whole evolves over time. The distribution over time for ${}^{2}D^{s}$ and five disciplines are shown in Figure 3, where dots represent mean values and bars indicate their 99% confidence intervals. Different periods are denoted as various colors in which darker colors represent more recent periods. Similar temporal trends are also found for RS and DIV.

It is obvious that publications from all studied disciplines in Social Sciences and Humanities are, at the aggregate level, becoming increasingly interdisciplinary over time, with a few distinctions for each. Linguistics was and continues to be the least interdisciplinary one among all five. Applied Psychology experienced the largest increase in the last 50 years. Although failing to gain as much growth as the others, Anthropology retains its leading position in interdisciplinarity. For the most recent 5 year period 2005–2009, Library Science is found to be the most interdisciplinary among all five.

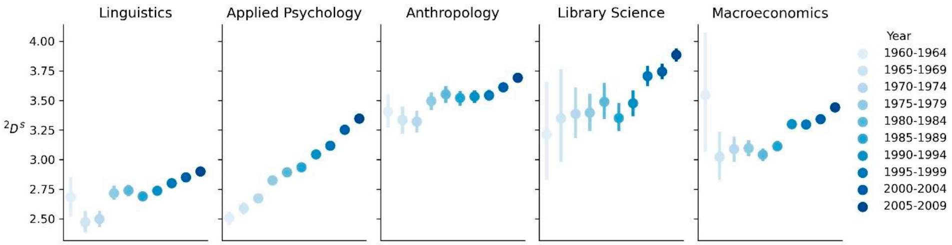


Figure 3. The evolution of integrated IDR measurements.

What should be additionally pointed out is that interdisciplinarity is not ever-increasing for all disciplines and all periods. In the mid-1960s, Linguistics and Anthropology exhibit a temporary drop in the level of interdisciplinarity on average, same for linguistics, Anthropology, and Library Science in 1985-1989. On the other hand, we can also observe certain synchronicity in change for certain periods, which suggests that although the developments in interdisciplinarity are realized by each individual discipline itself with their own pace or characteristics, they might be synchronously facilitated or hindered by a certain historical factor which could be academically related or otherwise.

The evolution of IDR outcome

Academic significance and visibility

To examine the potential impact of interdisciplinarity on academic significance and visibility, we determined citation impact with a 2- (C_2), 5- (C_5), and 10-year-long (C_{10}) citation window. The Spearman coefficient is calculated for each citation indicator and IDR indicator pair, as shown in Table 2. All correlations are statistically significant at the 0.1% level.

Table 2. Correlation between IDR measures, citation impact, and disruptiveness.

	Variety	***Balance***	***Disparity***	***RS***	***DIV***	$^2D^s$
C_2	.317	-.323	.094	.165	.229	.165
C_5	.395	-.406	.110	.209	.286	.209
C_{10}	.423	-.438	.117	.223	.306	.223
D_2	-.232	.277	-.048	-.092	-.147	-.092
D_5	-.276	.338	-.054	-.106	-.172	-.106
D_{10}	-.293	.363	-.055	-.109	-.180	-.109

Five out of six indicators turned out to be positively correlated with all three citation indicators, which indicates IDR's potential positive effect in terms of alleviating publications' academic significance and visibility. Balance, on the other hand, which measures the level of specialization in knowledge base, is negatively correlated with citation impact. This suggests that research may benefit from a more specialized perspective or knowledge base than focus-lacking ones. Variety is most positively correlated with citations among all.

Furthermore, the correlation becomes stronger as the citation window widens up which holds for all six IDR measurements. A possible interpretation of this observation is that the benefit interdisciplinarity might produce on citations requires time to form and may not be achieved overnight.

Disruption or consolidation

As explained in the “Methodology” section, the disruptiveness index (Wu et al., 2019) captures the proportion of net increase of new citing papers a publication invites to the local citation network. For instance, the publication will be regarded as disruptive with $D_t > 0$, if the number of its citing papers that also cite its references triumphs the number of citing papers that do not. In this case, the citing papers only acknowledge this paper’s contribution rather than its intellectual forebears which can be a sign of disruption. The counter case would be publications with $D_t < 0$, which means more of their citing papers also acknowledge their forebears, yielding consolidation.

This index has many advantages. Firstly, it possesses the potential capacity of recognizing scientific works that are commonly regarded as innovative and break-through in the field. A large-scale empirical study devised by Wu et al. (2019) found that papers that directly contribute to Nobel prizes tend to show high levels of disruptiveness while review articles are normally associated with low disruptiveness. What’s more, it can be more robust in resisting malicious citing behavior that is initiated by strategic considerations in one’s career rather than the quality of the work.

The correlations between disruptiveness and all of the IDR measurements are shown in Table 2. D_2, D_5, and D_{10} denote disruptiveness within a 2-year, 5-year, and 10-year time window, respectively. All correlations are significant at the 0.1% level. It seems that all IDR indicators except balance are negatively correlated with disruptiveness, which suggests interdisciplinary publications tend to consolidate fields or knowledge rather than disrupt them. Another interpretation is that research whose knowledge base is more diverse might eventually lead to scientific outcomes that are more likely to back up or build on prior knowledge. On the other hand, as captured by the correlation between balance and disruptiveness, a more specialized knowledge base could be beneficial in the production of disruptive research.

Similar to correlation with citation, IDR also seems to be more correlated with disruptiveness in a longer citation window. This suggests that disruption or consolidation one might produce in certain research domains will become more evident as time passes by.

An interesting question about the relationship between disruptiveness and interdisciplinarity is why the out-of-the-box thinking that interdisciplinary research strives for leads to consolidation while the more specialized publications turn out to be more capable of disrupting. A possible interpretation is that interdisciplinary publications tend to investigate emerging topics that are possibly heading towards the formation of a newborn field. The exploratory work these IDR research conducted can be conceived as creative or unorthodox in their creation but as consolidating when the topic emerges to discipline after a while. On the other hand, specialized scientific research is normally diving into specific topics that already have a solid foundation of previous knowledge and skill set. Limited room for consolidation can be found which yields more possibilities for disruptive outcomes.

Conclusions

This paper has examined the evolution of IDR in five fields from Social Sciences and Humanities (SSH), namely Anthropology, Applied Psychology, Linguistics, Library Science, and Macroeconomics over a 50-year-long time window. The IDR processes, embodied by disciplinary diversity in references, and outcomes, characterized by citation impact and disruptiveness, are studied and analyzed from a temporal perspective. We find that research from SSH is absorbing knowledge from an increasing number of increasingly “remote” disciplines, which leads to growth in the overall level of interdisciplinarity. Yet, the increasing level of specialization observed in the knowledge base is also significant which could be influenced by the formation or emergence of specific research topics. On the other hand, most of the IDR measurements and citations are found to exhibit positive correlations, which are strengthened as the citation window widens. Most of the IDR measurements and disruptiveness, however, are negatively correlated which becomes more evident with a longer time window.

There are many interesting implications based on our findings. The increasing trend of interdisciplinarity found in SSH further strengthens the evidence of the statement “science is becoming interdisciplinary”. This changing phenomenon calls for a necessity for us to reexamine the research paradigm in many disciplines and the evaluation system for various entities. Many research evaluations, for instance, peer review in journals, are conducted by specialized domain experts who might not have the relevant expertise for all interdisciplinary papers that integrates knowledge outside the discipline silos. The selection of a competent evaluation panel could be a somewhat challenging endeavor in the era of interdisciplinarity and should be handled seriously in a scientific way. Furthermore, we observe that the absolute value of the correlation between the number of citations and some interdisciplinarity indicators increases as the length of the citation time windows raises; and so does the degree of disruptiveness. This hints that interdisciplinary research needs time to show its potential, which is consistent with Wang et al., (2015). Previous studies have shown that interdisciplinary research is often encouraged by science policies, but they are insufficiently supported under current funding structures (Bromham et al., 2016). Thus, one implication for science policy decision-makers and funding providers is that assessing interdisciplinary research requires a longer time window.

Limitation and future work

This article has several limitations. For instance, only five disciplines in SSH are investigated which limits us to draw a comprehensive understanding for all SSH disciplines regarding the evolution of interdisciplinarity. In addition, the relationship between interdisciplinarity and research outcomes (citation and disruption) is analyzed in a simple way as a first step. A more sophisticated methodology that involves other confounding variables is required. In future studies, we will expand our current analysis to more disciplines in SSH and adopt advanced models to thoroughly uncover the interaction between interdisciplinarity and research outcomes.

Acknowledgments

This investigation has been made possible by the financial support of the Flemish Government to ECOOM, among others. The opinions in the paper are the authors’ and not necessarily those of the government.

References

Bromham, L., Dinnage, R., & Hua, X. (2016). Interdisciplinary research has consistently lower funding success. *Nature*, *534*(7609), 684–687. https://doi.org/10.1038/nature18315

Bu, Y., Li, M., Gu, W., & Huang, W. (2020). Topic diversity: A discipline scheme-free diversity measurement for journals. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24433

Carayol, N., & Thi, T. U. N. (2005). Why do academic scientists engage in interdisciplinary research? *Research Evaluation*, *14*(1), 70–79. https://doi.org/10.3152/147154405781776355

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. *American Sociological Review*, *80*(5), 875–908. https://doi.org/10.1177/0003122415601618

Frank, M. R., Wang, D., Cebrian, M., & Rahwan, I. (2019). The evolution of citation graphs in artificial intelligence research. *Nature Machine Intelligence*, *1*(2), 79–85. https://doi.org/10.1038/s42256-019-0024-5

Frodeman, R., & Mitcham, C. (2016). New Directions in Interdisciplinarity: Broad, Deep, and Critical. *Bulletin of Science, Technology & Society*. https://doi.org/10.1177/0270467607308284

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*. https://doi.org/10.1287/mnsc.2015.2366

Kong, X., Zhang, J., Zhang, D., Bu, Y., Ding, Y., & Xia, F. (2020). The Gene of Scientific Success. *ACM Transactions on Knowledge Discovery from Data*, *14*(4), 1–19. https://doi.org/10.1145/3385530

Larivière, V., & Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology*, *61*(1), 126–131. https://doi.org/10.1002/asi.21226

Larivière, V., Haustein, S., & Börner, K. (2015). Long-Distance Interdisciplinarity Leads to Higher Scientific Impact. *PLOS ONE*, *10*(3), e0122565. https://doi.org/10.1371/journal.pone.0122565

Leinster, T., & Cobbold, C. (2012). Measuring diversity: The importance of species similarity. *Ecology*, *93*, 477–489. https://doi.org/10.2307/23143936

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics*. https://doi.org/10.1016/j.joi.2018.12.006

Ma, Y., Mukherjee, S., & Uzzi, B. (2020). Mentorship and protégé success in STEM fields. *Proceedings of the National Academy of Sciences*, *117*(25), 14077–14083. https://doi.org/10.1073/pnas.1915516117

Mugabushaka, A.-M., Kyriakou, A., & Papazoglou, T. (2016). Bibliometric indicators of interdisciplinarity: The potential of the Leinster–Cobbold diversity indices to study disciplinary diversity. *Scientometrics*, *107*(2), 593–607. https://doi.org/10.1007/s11192-016-1865-x

Paszcza, B. (2016). *Comparison of Microsoft Academic (Graph) with Web of Science, Scopus and Google Scholar*. https://doi.org/10.13140/RG.2.2.21858.94405

Porter, A. L., Cohen, A. S., David Roessner, J., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, *72*(1), 117–147. https://doi.org/10.1007/s11192-007-1700-5

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, *82*(2), 263–287. https://doi.org/10.1007/s11192-009-0041-y

Rousseau, R., Zhang, L., & Hu, X. (2019). Knowledge Integration: Its Meaning and Measurement. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 69–94). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_3

Schummer, J. (2004). Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics*, *59*(3), 425–465. https://doi.org/10.1023/B:SCIE.0000018542.71314.38

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, *4*(15), 707–719. https://doi.org/10.1098/rsif.2007.0213

Szell, M., Ma, Y., & Sinatra, R. (2018). A Nobel opportunity for interdisciplinarity. *Nature Physics*, *14*(11), 1075–1078. https://doi.org/10.1038/s41567-018-0314-6

Thelwall, M. (2017). Microsoft Academic: A multidisciplinary comparison of citation counts with Scopus and Mendeley for 29 journals. *Journal of Informetrics*, *11*(4), 1201–1212. https://doi.org/10.1016/j.joi.2017.10.006

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*. https://doi.org/10.1016/j.joi.2010.06.004

Wang, J., Thijs, B., & Glänzel, W. (2015). Interdisciplinarity and Impact: Distinct Effects of Variety, Balance, and Disparity. *PLOS ONE*, *10*(5), e0127298. https://doi.org/10.1371/journal.pone.0127298

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396–413. https://doi.org/10.1162/qss_a_00021

Wang, Q., & Schneider, J. W. (2019). Consistency and validity of interdisciplinarity measures. *Quantitative Science Studies*, *1*(1), 239–263. https://doi.org/10.1162/qss_a_00011

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378–382. https://doi.org/10.1038/s41586-019-0941-9

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology*, *67*(5), 1257–1265. https://doi.org/10.1002/asi.23487

Zhang, L., Sun, B., Chinchilla-Rodríguez, Z., Chen, L., & Huang, Y. (2018). Interdisciplinarity and collaboration: On the relationship between disciplinary diversity in departmental affiliations and reference lists. *Scientometrics*, *117*(1), 271–291. https://doi.org/10.1007/s11192-018-2853-0

The Trend of Technological Convergence Based on Patent Analysis

Wenjing Zhu[1], Bohong Ma[2] and Lele Kang[*]

[1] *zhuwenjing@smail.nju.edu.cn*

[*]*lelekang @nju.edu.cn*
Nanjing University, School of Information Management, 305-1 Nanjing (China)

Abstract

With the development of modern technology, the boundaries between different technical fields are gradually blurring, and technological convergence has become an important research topic. Patent data is rich in information about technology and the development in various technical fields. Understanding the development trends of technology in patents is incredibly beneficial for future research and industrial development of advanced technology. This research analyzes the co-occurrence of technical fields at a large scale based on the International Patent Classification (IPC) numbers issued by the World Intellectual Property Organization (WIPO) and the industrial matching relationships in the technical fields. Subsequently, according to entropy theory, we investigate the technological convergence in 35 technological fields. After that, complementary convergence and substitution convergence among several fields are analyzed. It is found that the technological convergence of the chemical industry has been the most persistent in the past 20 years. The information technology industry has also grown rapidly in technological convergence in recent years. The results could help guide the direction of in-depth research within the entire industry and other prominent industries.

Introduction

As technology advances, the accelerating industrial innovation process has significantly increased the complexity and diversity of various technologies. In this progress, various technologies continuously integrate and interact, which causes a fusion phenomenon regarded as "technology convergence” (Griliches, 1990). In technology convergence, multiple technologies belonging to two or much more fields merge with others to develop a new function or product that has not existed previously (Hacklin, Marxt, & Fahrni, 2009). The significance of technology convergence is that technological revolutions are sometimes promoted by technology integration or reorganization of the basic knowledge of existing technologies in different fields, not just by the development of new technologies (Hacklin, 2008). For example, the integration of emerging electronic technologies and traditional mechanical technologies has provided us with mechatronics systems (Kodama, 2010).

As one of the most effective carriers of technical information, patents have the fastest updates, the broadest coverage, and the most comprehensive documentation of technical developments. They are the most reliable and reasonable indicators for understanding technical information development (Acs, Anselin, & Varga, 2002). Thus, we reviewed the literature on technological convergence based on patent analysis and noted the following issues. First of all, the definitions of "technological convergence" are diverse. Different scholars have some differences in their cognition and understanding of technological convergence.

Secondly, there is no agreement on the matching between technological patents and industrial classification in the literature. Lastly, there are various methods for measuring the degree of technological convergence. Most of the measurement methods are based on the International Patent Classification (IPC) code, and analyses rarely consider the industrial classification. Thus, this study is designed to investigate the technological convergence in different industries by taking these issues into consideration.

The rest of the paper is structured as follows. The following section starts with the definition and theories of technological convergence. After that, we introduce data and measurements of technological convergence. The third section presents a descriptive statistical analysis based on the measurement results. Lastly, we discuss the research results and give suggestions for future research directions.

Theoretical background

Definition of technological convergence

The concept of technological convergence was first proposed in 1963. Rosenberg (1963) found that there is very high similarity in terms of technology between different products produced by the machinery manufacturing industry and the metal processing industry. He interpreted this phenomenon as "technological convergence." This definition is the result of technological association. The academic community widely regards Rosenberg's work as the beginning of research in industrial integration. Since then, many scholars have cited the definition of technology convergence to describe the phenomenon of industrial technology convergence.

Lei (2000) believes that the boundaries between technology industries continue to weaken with the continuous emergence of innovative product concepts across multiple markets. Technology that belongs to one industry can have a profound impact on other industries or even multiple industries. From the perspective of industry evolution, Hacklin (2007) defines technological convergence as the potential impact of knowledge fusion on technological innovation capabilities. Knowledge spillover between industries promotes technological convergence, which leads to the emergence of new technologies. Here, technological convergence is regarded as the combination of various technical fields in patents.

Theories of technological convergence

The current patent-based technological convergence research has two major streams: patent citation analyses and patent classification analyses. The patent citation analysis method analyzes the citation network formed by the citation relationship that patent documents have with other patent documents or non-patent documents. The patent classification analysis method mainly analyzes the co-occurrence of different technical fields in patents to estimate the level of technological convergence. Leydesdorff, Kushnir, and Rafols (2014) used US patent information as sample data and established a patent similarity measurement model based on IPC codes and patent citation information. Criscuolo (2006) conducted an empirical analysis based on patent information provided by the patent databases of the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO) and found related families of patents.

The patent classification analysis method covers the primary technical fields and considers the effects of patent aging over time and technological differences. Generally, patent classification analysis methods can be divided into static and dynamic aspects. Static analyses of technological convergence consider interdisciplinary industry convergence in different technical fields to be necessary to continually search for and absorb relevant knowledge, as well as for emerging technologies generated by technological convergence to infiltrate into other industries. Therefore, the development directions of existing technologies and existing industries' business content have been continuously supplemented and improved.

Different technical fields share the same knowledge platforms for research and development. Lei (2000) and Stieglitz (2003) defined the mutually integrated technology fusion process as "Complementary Technological Convergence (CTC)" and defined the phenomenon of technologies in other fields eliminating current technologies as "Substitutability Technological Convergence (STC)." The technological convergence process can be seen as a combination of complementary technologies in multiple fields and a substitution of existing technologies.

From the dynamic perspective, technological convergence is not only a static phenomenon but also an evolutionary process (Yayavaram & Chen, 2015). The degree of convergence includes both the depth and width of technological convergence (Song, Almeida, & Wu, 2003). We can analyze the changes in the breadth and depth of technological convergence through the diversity of technological convergence. The current mainstream methods are mainly the Shannon Wiener index, Simpson index, and N index for dynamic analysis of technological convergence. All of them are based on the measurement methods of biological diversity and are formed by blending to analyze technical and economic fields. Stirling (2007) was the first to introduce biodiversity index analysis to a technical field. Gambardella and Torrisi (1998) used the Simpson index to conduct an empirical analysis of the electronic information industry's internal technological convergence.

Data and methodology

Data

This study analyzes patent data from the PATSTAT database. The patent classification system in the database is mainly based on the International Patent Classification IPC code. The IPC system is an accurate, useful, and easy-to-use tool for classifying and searching patent applications, authorized patent descriptions, utility models, and similar technical documents. It divides technical fields into eight divisions (A-H) with about 75,000 subclasses. Every subclass is represented by a language-independent classification number, which is composed of letters and Arabic numerals. IPC consists of multiple levels, and several dots indicate the grouping level. The more dots, the lower the grouping level is.

Many scholars around the world have explored the consistency between the IPC system of patents and the classification of industrial technology. One famous example is the concordance table that records the relationship between IPC and International Standard Industrial Classification (ISIC). In this study, we use the IPC and technology concordance table from the World Intellectual Property Organization (WIPO) because the ISI-OST-INPI classification method proposed by Schmoch (2008) combines theoretical analysis and empirical research. Furthermore, it entirely takes into account the differences in industrial classifications in different countries. Therefore, the types of industrial technology designs can be applied to international comparisons with broad applicability and strong persuasiveness.

First of all, we selected patent data with filing years between 2000 and 2019 from the PATSTAT database. Next, we matched IPC codes and technical fields according to the ISI-OST-INPI classification. Finally, we performed de-duplication operations to ensure that the one patent and one technical field only match once. We collected 17,755,027 patents through this process, and there were 23,941,243 data records in total over the period of 2000-2019. By filtering the data, we found that the patent data for 2019 is somewhat incomplete, so we excluded the data for 2019 and only considered the filing years between 2000 and 2018, leaving a total of 23,770,593 data records.

Methodology

The current measurement method for patent technological convergence mainly uses patent indicators for evaluation. Based on the literature, we have compiled the results of various index measurement methods in academia, as shown in Table 1 (Li & Shu, 2017). Considering the large amount of information in our selected database, the patents involved are numerous and complex.

Table 1. Measurement indicators of technological convergence (Li & Shu, 2017).

	Definition	Formula
Macro-level	Technological convergence: The ratio of the number of cross-patents (multi-technology classification) to the total number of patents (full field)	$Technological\ convergence = \frac{Numbers\ of\ cross\ patents}{Numbers\ of\ total\ patents}$
	Technological coverage: Using Shannon Entropy to calculate the technological convergence breadth in a particular field, the F value changes with time. The greater the F value, the greater the convergence breadth of the technological field.	$Fi = -\sum_{k} P_{ik} lnP_{ik}$ Fi: The entropy of field i; P_{ik}: Percentage of field k in field i.
Meso-level	Technological cross-convergence coverage: The coverage of cross-technological convergence.	$W_i = \frac{Ui}{UA}(i \in A)$
	Technological cross-convergence intensity: The intensity of cross-technological convergence.	$Ii = \frac{Pi/Ui}{PA/UA}(i \in A)$, Cross technology i belongs to industry A, Ui and Pi represent the number of USPCs and patents included in cross technology i; UA and PA are the number of USPCs and patents included in industry A.
	Convergence intensity: The ratio of the number of cross patents to the minimum number of patents in the two different source technologies.	$CIA - B = N_{P_{A-B}}/Min\ (N_{PA}, N_{PB})$ $CIA - B$: The convergence intensity of field A and field B; $N_{P_{A-B}}$: Numbers of patents in field A and field B.
	Convergence coverage: The number of patented technologies involved in the same category is used to measure the degree of diffusion in the convergence process of a specific technical field.	$CC = \frac{C}{M \times N}$
Micro-level	Originality: Measure the originality of a technology by the presence of patents outside its classification.	$OI = 1 - \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2$, k represents the number of different technical fields to which the patents cited in the observed patent citations belong; N_i represents the number of cited patents belonging to technical field I; N represents the total number of cited patents.
	Generality: Measure of the generality of the technology based on the citation volume and the classification distribution of the cited patents.	$GI = 1 - \sum_{i=1}^{k} \left(\frac{N_i}{N}\right)^2$, k represents the number of different technical fields to which the patent citing the observation (citing patent) belongs; N_i represents the number of citing patent belonging to technical field; N represents the total number of citing patents.

Therefore, according to the various index measurement methods mentioned above, this study firstly adopted Shannon entropy algorithm to measure the degree of convergence of technical fields and to find the fields that are more actively convergent for more in-depth research. We measured the complementary and substitutability technological convergence in four steps based on the measurement method by Dibiaggio, Nasiriyar, and Nesta (2014).

In step 1, we used four or five years as a time interval and divided the period of 2000-2018 into four stages. We then constructed a 35*35 co-occurrence matrix of technical fields according to the time interval. If two different technical field codes appeared once in the same patent, a technical co-occurrence was considered to have appeared in the two fields. We chose 5*5 technological fields to show in Table 2.

Table 2. Technology co-occurrence matrix (sample from 2000-2004).

	Analysis of biological materials	*Audio-visual technology*	*Basic communication processes*	*Basic materials chemistry*	*Biotechnology*
Analysis of biological materials	34401	256	22	1041	20649
Audio-visual technology	256	191490	8210	2634	228
Basic communication processes	22	8210	43205	7	29
Basic materials chemistry	1041	2634	7	94370	4607
Biotechnology	20649	228	29	4607	81089

In step 2, taking the technology co-occurrence matrix as a benchmark, we found the Shannon entropy based on the following equation:

$$Entropy_i = -\sum_j p_{\frac{j}{i}} \, log_2 p_{\frac{j}{i}} \tag{1}$$

where *p* represents the ratio of the number of patents in technical fields *i* and *j* to the number of patents in technical field *i*. In step 3, we chose the top-five technical fields with the highest entropy, counted the co-occurrence frequencies between different technical fields, and estimated the mean value and standard deviation based on equations (2) and (3).

$$\mu_{ij} = E(X_{ij} = x) = \frac{O_i O_j}{K} \tag{2}$$

$$\sigma_{ij}^2 = \mu_{ij}\left(\frac{K - O_i}{K}\right)\left(\frac{K - O_j}{K - 1}\right) \tag{3}$$

O_i represents the number of all patents involved in technical field *i*, and O_j represents the number of all patents involved in technical field *j*. *K* stands for a summary of all patent applications.

In step 4, according to the basic calculation equations above, the CTC index was calculated using equation (4):

$$CTC = \frac{C_{ij} - \mu_{ij}}{\sigma_{ij}} \tag{4}$$

where C_{ij} is the total number of patents involving patents *i* and *j* at the same time. In step 5, we use a cosine function to measure the convergence of substitutability technology using the STC index:

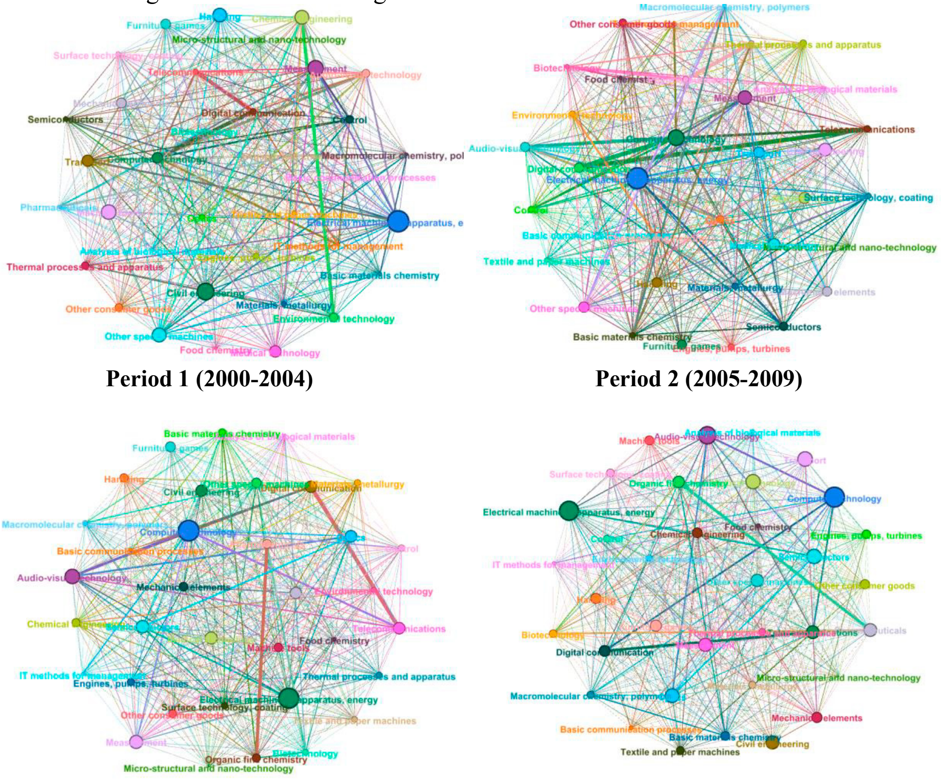
$$STC = \frac{\sum_{n=1}^{k} C_{in} C_{jn}}{\sqrt{\sum_{n=1}^{k} C_{in}^{2}} \sqrt{\sum_{n=1}^{k} C_{jn}^{2}}} \tag{5}$$

where *k* represents the total amount of patents applied for, C_{in} represents the frequency of the technical element with code *i* and *n* other technical fields appearing in the same patent, and C_{jn} represents how often the technical field with code *j* and *n* other technologies appears in the same patent.

Empirical analysis

Network visualization

Graphs can represent most scientometric data-based networks. They are applied to collaboration networks as well as to co-citation networks (Glanzel, 2012). Network visualization provides the most direct display of the cross mode in technological convergence in this study. We introduce a node analysis at the technical-field level. According to the co-occurrence matrix results of four five-year windows, the number of patents in each field is used as point data, and the number of co-occurrences between fields is used as side data. The node circle size and the thickness of the line connecting the two nodes respectively represent the strength of the domain attribute and the degree of co-occurrence. The four final network visualization diagrams are shown in Figure 1.



Period 1 (2000-2004)

Period 2 (2005-2009)

Period 3 (2010-2014)

Period 4 (2015-2018)

Figure 1. Network visualization of technical fields by periods.

Comparing the network visualization diagrams of technical fields in the four periods, we can find that the main fields involved in the patent applications have transformed from the *computer technology* field to the field *of electrical machinery, apparatuses, and energy* since around 2010. The largest node in Period 1 and Period 2 is *computer technology*, followed by *electrical machinery, apparatuses, and energy*, and the third one is *audio-visual technology*. In Period 3 and Period 4, the size of the *computer technology* field is gradually decreasing, while the node of *electrical machinery, apparatuses, and energy* becomes the largest one, followed by *audio-visual technology*.

The technical fields with more patents do not represent a higher impact on the other fields based on the visualization. Among the four periods, the two areas with the most co-occurrences are the *organic fine chemistry* field and *pharmaceuticals* field in Period 1, the *digital communication* field and *telecommunications* field in Period 2 and Period 3, and the *organic fine chemistry* field and *digital communication* field in Period 4. However, these technical fields with the most co-occurrences are not the fields with the largest nodes. Thus, the co-occurrence of technical fields is not necessarily related to the number of patents in the technical field.

STC and CTC

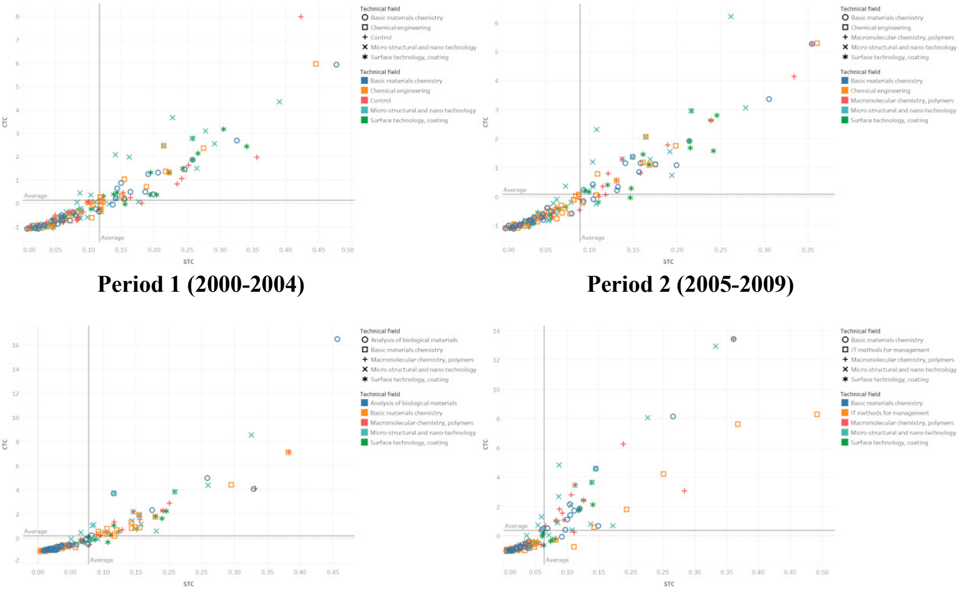
STC and CTC also show technological convergence. Table 3 shows the empirical results of Shannon entropy's estimation, in which there are five technical fields with enormous entropy values in each time window. Using this as a benchmark, we calculated the STC and CTC of each field with the other 34 fields.

Table 3. The highest entropy of technical fields (by period).

Period	*Technical field*	*Entropy*
2000-2004	Micro-structural and nano-technology	6.6356036710
	Surface technology, coating	5.7884236280
	Chemical engineering	5.6370176984
	Basic materials chemistry	5.4449187914
	Control	5.0690933392
2005-2009	Micro-structural and nano-technology	5.8079235452
	Surface technology, coating	5.0940984975
	Basic materials chemistry	4.7347813344
	Chemical engineering	4.5223951158
	Macromolecular chemistry polymers	4.5163670576
2010-2014	Micro-structural and nano-technology	5.6658510134
	Surface technology, coating	4.5690331727
	Macromolecular chemistry polymers	4.5425173526
	Basic materials chemistry	4.4230508371
	Analysis of biological materials	3.8200670523
2015-2018	Micro-structural and nano-technology	3.3441336768
	Macromolecular chemistry polymers	2.8603060175
	Basic materials chemistry	2.5983972381
	IT methods for management	2.3276341583
	Surface technology, coating	2.2365707230

To better understand the evolution of individual technical fields within the patent network, we constructed a two-dimensional quadrant with a substitutability technology convergence axis and complementary technology convergence. This matrix was depicted for the four periods with the STC indicating the compatibility between related technologies and the CTC for the differences. The first quadrant is defined as the convergence of high-substitution and high-complementarity technologies. This phenomenon is the mature stage of the convergence, where both substitution and complementarity have reached a certain level.

The second quadrant's technical fields have a higher CTC and a lower STC, which means that most of the two technologies in this quadrant are gradually integrated through reorganization and creation in the same environment. In the third quadrant, STC and CTC values between the two different technical fields are both low. It is indicated that the two technologies in this quadrant are relatively independent, and the degree of convergence of various technologies between industries is weak. Finally, in the fourth quadrant, the STC value is high, whereas the CTC value is low. This quadrant's technical fields mainly improve the reliability of new products and accelerate market acceptance through the compatible docking of similar technical elements.



Period 1 (2000-2004) **Period 2 (2005-2009)**

Period 3 (2010-2014) **Period 4 (2015-2019)**

Figure 2. Scatter plot of technical fields by periods.

Our study discovered that the values of STC and CTC have a certain degree of correlation. Figure 2 shows that scattered points are distributed in the third quadrant and are scattered around the diagonal with a slope of 1. This means that the process of technological convergence and industry evolution requires complementary convergence to promote industrial innovation and substitutability convergence to promote industrial evolution.

With subdivision into four periods for comparison, we can see that there has been a significant increase in both the abscissa and the ordinate, which confirms the increase in technological convergence between different industries. In addition, according to the distribution of the points in the first quadrant and the entropy ranking of the four periods in Table 1, we can see that although the entropy value of the *microstructure and nano-technology* field has always been the highest, the proportion of points of the field in the first quadrant is being pursued by other fields. The range is gradually shrinking from period 1 to period 4. This shows that technological convergence is no longer limited to several fields but is spreading to various industries, and industrial boundaries are constantly blurring. We list the top five industry combinations with the highest STC and CTC in the four periods in Table 4 for further analysis and comparison.

Table 4. The highest STC and CTC of technical fields from 2000 to 2004.

Period 1	*Technical field*	*Sub-field*
STC	Basic materials chemistry	Macromolecular chemistry, polymers
	Chemical engineering	Environmental technology
	Control	IT methods for management
	Micro-structural and nano-technology	Semiconductors
	Control	Computer technology
CTC	Micro-structural and nano-technology	IT methods for management
	Surface technology, coating	Environmental technology
	Basic material chemistry	Macromolecular chemistry, polymers
	Chemical engineering	Semiconductors
	Macromolecular chemistry, polymers	Materials, metallurgy

During the period of 2000-2004, the high STC values within the chemical industry and environmental industry are closely related to promoting the environmental industry as a "green industry" or "sunrise industry" in the early 21st century. Although the semiconductor industry has reached the end of its heyday, it still has strong technological convergence with the *micro-structural and nano-technology* field. The *IT industry* has gradually begun to rise and has closely converged with the *control* industry, which belongs to the instruments category.

Table 5. The highest STC and CTC of technical fields from 2005 to 2009.

Period 2	*Technical field*	*Sub-field*
STC	Chemical engineering	Environmental technology
	Basic materials chemistry	Macromolecular chemistry, polymers
	Macromolecular chemistry, polymers	Other special machines
	Basic materials chemistry	Organic fine chemistry
	Micro-structural and nano-technology	Semiconductors
CTC	Micro-structural and nano-technology	Materials, metallurgy
	Chemical engineering	Environmental technology
	Basic material chemistry	Macromolecular chemistry, polymers
	Macromolecular chemistry, polymers	Other special machines
	Basic materials chemistry	Organic fine chemistry

In the period of 2005-2009, regardless of STC and CTC values, the top five fields are almost all chemical-related technologies. It can be seen that the chemical industry in this period is in a stage of transition between old and new. Compared with the first period, *raw materials chemistry* fields are distributed less than *chemical engineering* fields within the first quadrant. This reflects the development trend of the chemical industry to a certain extent. Also, combined with the Shannon entropy in Table 2, *control* technologies are no longer a field of high technological convergence during this period. The changes in industry development trends can be predicted from this.

Table 6. The highest STC and CTC of technical fields from 2010 to 2014.

Period 3	*Technical field*	*Sub-field*
STC	Analysis of biological materials	Biotechnology
	Basic materials chemistry	Macromolecular chemistry, polymers
	Macromolecular chemistry, polymers	Other special machines
	Analysis of biological materials	Measurement
	Micro-structural and nano-technology	Materials, metallurgy
CTC	Analysis of biological materials	Biotechnology
	Micro-structural and nano-technology	Materials, metallurgy
	Basic material chemistry	Macromolecular chemistry, polymers
	Analysis of biological materials	Pharmaceuticals
	Basic materials chemistry	Organic fine chemistry

For the period of 2010-2014, an incredibly unique phenomenon is the distribution *of biological materials*, as shown in Figure 2 and Table 6. The convergence relationship between the *analysis of biological materials* and *biotechnology* is quite prominent. Both CTC and STC are much higher than for other technical fields. However, in the top-five technological fields in terms of entropy, there is less distribution of *analysis of biological materials* in the first quadrant than for the other four fields. However, CTC and STC values are higher than for other fields, such as the convergence with *measurement* and *pharmaceuticals*. The "Pareto principle" has been fully embodied in the *analysis of the biological materials* field.

Table 7. The highest STC and CTC of technical fields from 2015 to 2019.

Period 4	*Technical field*	*Sub-field*
STC	IT methods for management	Computer technology
	IT methods for management	Digital communication
	Basic materials chemistry	Macromolecular chemistry, polymers
	Micro-structural and nano-technology	Materials, metallurgy
	Macromolecular chemistry, polymers	Other special machines
CTC	Basic materials chemistry	Macromolecular chemistry, polymers
	Micro-structural and nano-technology	Materials, metallurgy
	IT methods for management	Computer technology
	Basic materials chemistry	Organic fine chemistry
	Micro-structural and nano-technology	Semiconductors

During the period of 2015-2018, the technology convergence in the *IT methods for management* dramatically increased. The value of STC for *computer technology* and *digital communication* shows that the ICT industry is developing at its peak. We believe that the development of miniaturization technology has led to the convergence of alternative products between handheld computers and mobile devices. While companies continue to use innovative opportunities advantages, they have also promoted this industry's continuous development. The internal technology convergence of the chemical industry and the electrical engineering industry are still strong.

Combined with Table 3, it is easy to see that the *micro-structural and nano-technology* field has always had the highest Shannon entropy during the past 20 years. Analysis of this domain alone reveals that although the entropy has been highest, the entropy of this field has been decreasing with time. We believe that this is related to the industry-wide diffusion of technology convergence, which is no longer concentrated in individual fields due to the expanding scope of technology convergence. At the same time, the values of STC and CTC in the *micro-structural and nano-technology* domain do not exhibit particularly high values, which we believe validates the theory of technology convergence diffusion.

Conclusion

This study has investigated technological convergence based on patent analysis by developing and applying concepts of Shannon entropy, complementary technological convergence, and substitutability technological convergence. We believe that the quadrants of STC and CTC indicate the unique patterns and technological convergence trends. Furthermore, we have found the evolutionary patterns of technological convergence and provided a valuable blueprint for future research.

Our study is limited in terms of its measure of technological convergence. Patent data is important for identifying the intersection of technical fields. We only used the IPC patent analysis method to measure the intersection of technical fields. Thus, future research could introduce a patent citation analysis method into the analysis framework, which could provide a more accurate measurement tool to comprehensively examine characteristics of cross-field technological convergence.

Acknowledgments

The authors thank Ruichun Peng for his wonderful figures. We also thank Chang Zhang for the help and advice provided. Funding was obtained from the National Natural Science Foundation of China under Grant 72072087, which is also gratefully acknowledged.

References

Acs, Z. J., Anselin, L. & Varga, A. (2002). Patents and innovation counts as measures of regional production of new knowledge. *Research Policy,* 31(7), 1069-1085.

Comanor, W. S. & Scherer, F. M. (1969). Patent Statistics as a measure of technical change. *Journal of Political Economy,* 77(3), 392-398.

Criscuolo, P. (2006). The 'home advantage' effect and patent families. A comparison of OECD triadic patents, the USPTO and the EPO. *Scientometrics,* 66(1), 23-41.

Dibiaggio, L., Nasiriyar, M. & Nesta, L. (2014). Substitutability and complementarity of technological knowledge and the inventive performance of semiconductor companies. *Research Policy,* 43(9), 1582-1593.

Gambardella, A. & Torrisi, S. (1998). Does technological convergence imply convergence in markets? Evidence from the electronics industry. *Research Policy,* 27*(5)*, 445-463.

Glänzel, W. (2012). The role of core documents in bibliometric network analysis and their relation with h-type indices. *Scientometrics,* 93(1), 113-123.

Griliches, Z. (1990). Patent Statistics as Economic Indicator - A Survey. *Journal of Economic Literature,* 28(4), 1661-1707.

Hacklin, F. (2007). Management of Convergence in Innovation. *Contributions to Management Science,* 2010(42), 1014-1021.

Hacklin, F. (2008). *Management of Convergence in Innovation: Strategies and Capabilities for Value Creation Beyond Blurring Industry Boundaries.* Physica-Verlag GmbH.

Hacklin, F., Marxt, C. & Fahrni, F. (2009). Coevolutionary cycles of convergence: An extrapolation from the ICT industry. *Technological Forecasting and Social Change,* 76(6), 723-736.

Iansiti, M. (1997). *Technology Integration: Making Critical Choices in a Turbulent World.* Harvard Business School Press.

Kodama, F. (2010). Emerging patterns of innovation: sources of Japan's teleological edge. *R & D Management,* 26(2), 179-181.

Lee, K. R. (2007). Patterns and processes of contemporary technology fusion: The case of intelligent Robots. *Asian Journal of Technology Innovation*, 15(2), 45-65.

Lei, D. T. (2000). Industry evolution and competence development: the imperatives of technological convergence. *International Journal of Technology Management,* 19(7-8), 699-738.

Leydesdorff, L., Kushnir, D. & Rafols, I. (2014). Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC). *Scientometrics,* 98(3), 1583-1599.

Li, S. & Shu, F. (2017). Review of data analysis methods in measuring technology fusion and trend. *Data Analysis and Knowledge Discovery,* 1(7), 2-12.

Narin, F. & Hamilton, K. S. (1996). Bibliometric performance measures. *Scientometrics,* 36(3), 293-310.

Rosenberg, N. (1963). Technological-change in the machine-tool industry, 1840-1910. *Journal of Economic History,* 23(4), 414-443.

Schmoch, U. (2008). Concept of a technology classification for country comparisons. *Final Report to the World Intellectual Property Organisation*.

Song, J., Almeida, P. & Wu, G. (2003). Learning–by–Hiring: When is mobility more likely to facilitate interfirm knowledge transfer? *Management Science*, 49(4), 351-365.

Stieglitz, N. (2003). Digital dynamics and types of industry convergence: The evolution of the handheld computers market. In: *The industrial dynamics of the new digital economy,* 179-208.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface,* 4(15), 707-719.

Yayavaram, S. & Chen, W. R. (2015). Changes in Firm Knowledge coupling and firm innovation performance: The moderating role of technological complexity. *Strategic Management Journal,* 36(3), 377-396.

Impact of Model Settings on the Text-based Rao Diversity Index

Andrea Zielinski[1]

[1] *andrea.zielinski@isi.fraunhofer.de*
Fraunhofer Institute for Systems and Innovation Research ISI, Breslauer Strasse 48, 76139 Karlsruhe (Germany)

Abstract

Topic models such as Latent Dirichlet Allocation (LDA) have been proved to be effective tools to discover latent topics in text collections in a data-driven way. These topics can be further utilized to investigate academic disciplines in terms of *interdisciplinarity* by means of indicators that reflect the diversity of the scientific output. This study provides a systematic analysis of model parameters that affect the diversity scores which are computed directly from the output of the LDA model.

We present an empirical study on a real data set, upon which we quantify the diversity of the research within several departments of Fraunhofer (FH) and Max Planck Society (MPG) by means of scientific abstracts published in Scopus between 2008 and 2018. Our experiments show that parameter variations, i.e. the choice of the number of topics, hyper-parameters, and size and balance of the underlying data used for training the model, have a strong effect on the LDA-based Rao metrics. In particular, we could observe sharp fluctuations of the Rao index when varying over the number of topics. Due to its instability, it might not be a useful indicator of *interdisciplinary.*

Introduction

Interdisciplinary research (IDR) is a mode of research that integrates information, data, techniques, tools, perspectives, concepts, and theories from two or more scientific disciplines. According to innovation theory, research addressing social and economic needs is often beyond the scope of a single discipline and therefore policy-makers often promote IDR (see National Academies (2005)[1].

The most frequently used method to operationalize the concept of IDR is by means of the multi-dimensional Rao-Stirling indicator (Stirling, 2007) which contains three different dimensions: (1) variety: number of distinctive categories; (2) balance: evenness of distribution; and finally (3) disparity: degree to which the categories are different. In bibliometrics, the diversity score considers the number of publications in a scientific category and/or the percentage of references to documents into other scientific disciplines and relies on the metadata of scientific publications (Leydesdorff & Rafols, 2009).

According to Cassi et al. (2017), Rao is a relevant indicator at the scale of a research institution and can be adopted for comparing institutions' interdisciplinary practices but requires a proper delineation into research fields. Even though major publishers such as Elsevier provide a categorization scheme designed to define a scientific discipline, e.g. the ASJC codes in Scopus, the classification of articles is often too imprecise and course-grained for measuring interdisciplinarity, since articles are assigned to subject categories associated with the journal rather than the article (Zhang et al., 2016).

In contrast, clustering approaches based on machine learning allow to produce more fine-grained, faceted topics of the research literature. In addition, they are able to classify scientific knowledge into novel categories without the need to resort to human-defined subject categories that might be outdated (Suominen and Toivanen, 2016). In particular probabilistic topic models such as LDA (Blei et al., 2003, 2010) have been applied to the task of mapping research into fields of science (Yau et al., 2014).

Topic models have also been used to capture the notion of *interdisciplinarity* of research institutions, either based on scientific publications (Paul and Girju, 2009; Nanni et al., 2016) or research awards (Nichols, 2014; Talley et al., 2011).

When dealing with large datasets, employing ML algorithms that are able to calculate indicators in an unsupervised fashion are particularly attractive. An appealing work in this direction is provided by Bache et al. (2013) and Wang et al. (2014) who have re-interpreted the Rao Stirling indicator on the basis of topic modeling, relying exclusively on textual features. The authors conduct experiments on synthetic as well as real data sets (using abstracts, full papers, or grants) that suggest that also the text-based implementation of Rao's index correlates with human judgements.

Topic models are popular because of their data-driven nature that seeks to find emerging clusters of scientific disciplines automatically. Furthermore, they are multi-mixture models where a document may contain several topics. Yet, it is well known that purely unsupervised models such as LDA often result in topics that do not fit the needs of a specific application, i.e. they do not necessarily align with an established subject domain classification schema. Moreover, hyper-parameter setting is important to produce high quality topics (Syed and Spruit, 2018; Chang et al., 2009). According to Tang et al. (2014), LDA's performance depends mainly on the factors a) number of topics, b) the Dirichlet (hyper)parameters, c) number of documents, and d) the length of individual documents.

One of the most crucial factors is the number of topics: Standard LDA requires that a good estimate of the number is known to avoid over-/underfitting of the data. By design, LDA topic models often make use of the sparse Dirichlet priors such that each document contains only a small number of topics and each topic uses only a small set of words frequently. Yet, setting these hyper-parameters has an impact on the document-topic and topic-word distribution and leaves room for variation.

This paper seeks to investigate in a pilot study in how much the LDA-based Rao measure is sensitive to parameter settings and if it can be used as a reliable indicator to automatically calculate a diversity ranking according to an institute's research output, i.e. based on abstract and title as listed in Scopus.

The rest of this article is organized as follows. First, we briefly discuss related work. In the second section, we summarize the definition of the Rao-based disciplinarity indicator, and discuss the topic-specific calculation of the metrics on the basis of LDA. Then, we briefly introduce the data used for the empirical analyses. Subsequently, we present the experimental results on the publication output of two research institutes. Finally, we conclude the article and state future directions.

Related Work

Establishing methods for defining and measuring interdisciplinarity is central and intensively studied within bibliometrics (Wagner et al., 2011). The main goal of the task is to automatically define reliable indicators that are efficient to calculate, predictive, and robust regarding data errors (Guo et al., 2009).

A well established indicator has been set up by Rao (1982) and Stirling (2007), i.e. the Rao-Stirling diversity, which considers variety (number of distinct categories), balance (evenness of the distribution), and disparity (distances or similarities between categories). Accordingly, variety is defined as the number of subject categories assigned to the papers' references and takes values between one and the number of subject categories[ii], balance is a function of assignments across categories and ranges between zero and one[iii], and disparity is the complement of similarity and computed pairwise between the referenced subject categories. Its value also ranges between zero and one[iv]. Yet, the bibliometric operationalization of diversity is actively discussed in the research community (Leydesdorff, 2018; Leydesdorff, 2019). Based on a case study on Web of Science data, Wang and Schneider (2020) found that many measures are inconsistent. This also holds for the Rao-Stirling indicator which has recently been criticized for its low discriminatory power (Zhou et al., 2012).

Starting from the pioneering works by Hall et al. (2008), Paul and Girju (2009), Griffiths and Steyvers (2004), among others, models of diversity have also spread in the area of computational linguistics, especially in connection with topic modelling. These approaches all rely on accepted subject classifications from journals or conference proceedings. Paul and Girju (2009) assess the interdisciplinary nature of distinct research fields based on their topic overlap. Document collections featuring different research fields are compared via their mean topic vectors using cosine similarity. Nichols (2014) apply LDA topic modelling to analyze research awards issued at the National Science Foundation (NSF), inducing 1,000 latent topics from 170,000 project award descriptions. The institutional structure serves as a proxy for research disciplines and topics are assigned to the discipline in which they occur most frequently. The author observed a high variation in the topic frequency over years, while the aggregation of the topics into disciplines accounted for the temporal stability and resulted in a relatively constant score between 0.11 and 0.125.

In contrast, Bache et al. (2013) define the Rao measure entirely on the LDA output, without mapping topics to pre-defined classes that reflect specific scientific disciplines, and without verifying the nature of the topics. In their work, the Rao index is derived in a fully data-driven way and computed on the level of a document over the LDA document-topic and word-topic matrices. The authors conduct various experiments on PubMed Open Access, NSF Grant Awards, and the ACL Anthology and build a topic model for each corpus, varying over the number of topics (K = 10, 30, 100 and 300) and keeping the hyper-parameters fixed, in order to compute the Rao diversity scores. The authors state that the topic-based Rao diversity measure outperforms alternative approaches like entropy in a classification task on pseudo documents. The authors hypothesize that the method would be invariant to the number of topics in the model. Wang et al. (2014) use the same approach as Bache et al. (2013), however, their LDA model is induced from a corpus that considers a paper's references and citations. The authors propose a discounting weight on the balance attribute as part of the diversity score.

Furthermore, a variety of LDA models has been proposed to address certain limitations of LDA and give better performance, for instance, when it comes to detecting rare topics in an imbalanced collection (Jagarlamudi et al., 2012) or short text (Newman et al., 2011; Quan et al., 2015). Incorporating meta-information directly into the generative process of topic models can improve modelling accuracy and topic quality. Various authors have used document labels as a priori information to infer the underlying topic distributions, using training data with known labels in a semi-supervised setting (Ramage et al., 2010). Topic models have been often used in combination with partially supervised methods (Chuang et al., 2012). It has been shown that document regularization yields improved model performance, however requires reliable labeled data (Zhao et al., 2017).

Rao Stirling Diversity Measures based on LDA

The classic Rao Stirling diversity index has been widely used to measure diversity and interdisciplinarity (e.g. Porter & Rafols, 2009; Wang et al., 2015). In this section, we will discuss the three different dimensions of diversity i.e. variety, balance, and disparity.

Variety

Instead of subject categories, the thematic diversity can be related to the number of distinct topics K. A characteristic of latent topics generated by LDA, however, is that every topic is in principle present in every document, with a non-zero proportion. A rough estimate is that a large number of topics is needed to account for small scientific communities. Current approaches set the number of topics between K=300 (Griffiths et al., 2004) and K=1,000 (Nichols, 2014) to cover the whole scientific landscape. Griffiths et al. (2004) determine the number of topics based on the log-likelihood of the data, while Nichols (2014) set the number of topics according to the number of research divisions at NSF. In practice, a higher number of topics will necessarily result in a larger variety. This issue is crucial because the optimal number of topics in a corpus is unknown and based on a heuristic choice.

Balance

Generally, a more balanced document-topic distribution results in a higher thematic diversity estimate. The balance component as part of the Stirling Index can be calculated as follows:

$$\sum_{i=1}^{K}\sum_{j=1\,(i\neq j)}^{K} P(i|d)\,P(j|d) \qquad \forall\ \mathrm{d}: \min^{\mathrm{T}\,(\mathrm{T}-1)} \leq\ \mathrm{B} \leq \max^{\mathrm{K}\,(\mathrm{K}-1)}$$

where *P(i|d)* is the probability of topic i in a paper d and individual pair scores take small values in the range of *[min*: $\sim 10^{-6}$,*max*: $\sim$0.25]. Regarding the distribution of papers into scientific categories, it is likely that any database that seeks to monitor scientific research will consist of long-tailed, imbalanced data that is prevalent in any real-world setting. In order to deal with the issue of imbalanced data, it is necessary to have a good estimate of the scalar concentration parameter α that governs the shape of the document-topic distribution. Setting α to a value close to zero will result in a distribution where the probability mass is concentrated on a smaller set of topics. Moreover, an asymmetric learns a non-uniform prior, assuming that certain topics might be more prominent in the collection. Thus, some topics may be the majority topic in a larger share of documents in the corpus overall and make up more of the total corpus. As an alternative, proper sampling methods that re-balance the data can help to mitigate the problem.

Disparity

Topic similarity metrics can be applied to estimate the (dis)similarity $\delta(i,\ j)$ between topics i and j, and are generally computed from the topics' word probability distributions. A systematic evaluation of different topic similarity measures for pairs of topics generated by LDA has been conducted by Aletras et al. (2014) and Wang et al. (2019), comparing which measure aligns best with human judgements. Their experiments show that intrinsic coherence scores like Jensen-Shannon, Hellinger, Jaccard Distance and cosine similarity applied on the original dataset are generally inferior to extrinsic metrics that make use of external data. However, it is crucial that the external datasets fit well to the domain of the data used to build the topic model. In the setting of Aletras et al. (2014), co-occurrences of words were drawn from Wikipedia, while Wang et al. (2019) use word embeddings, which have been specifically trained on Twitter data. Since external data that covers the immense variety of scholarly topics is not readily available, we use intrinsic measures to compute topic similarity. An alternative approach proposed in Bache et al. (2013) is to make use of the document-topic matrix in order to calculate the probability or cosine distance of distinct topics that co-occur in documents. The motivation for this approach is that topic distributions tend to be distinct by definition. We refrain from this approach, because standard LDA is unable to model relations among topics due to its use of a single Dirichlet distribution, and thus it is not possible to detect correlations amongst topics directly. In order to transform the similarity matrix between topics i and j into a dissimilarity matrix, a frequently applied solution is 1- δ(i, j) and 1/δ(i, j). Based on prior studies, we choose the metrics listed in Table 1 for our evaluation study. The topic distance also indicates how well the topics are separated which is a sign for a high quality LDA model. In order to produce topics that are distinct from each other, a symmetric prior of the topic-word distribution is generally preferred, and the β hyper-parameter needs to be set to values ranging between 0.1 and 0.01, so that the topic vectors concentrate on fewer words (Wallach et al., 2009).

Table 1. Topic Similarity Measures based on the Topic-Word Matrix.

Metrics	***Measure***	***Author***
Divergence-based metrics	*JS Divergence*	Hall et al. (2008)
Coefficient-based metrics	*Jaccard*	Ramage et al. (2009)
Distance-based metrics	*Hellinger Distance*	Aletras et al. (2014)
	Cosine	Wang et al. (2019)

Summary of Diversity Measures

We apply the Rao-Stirling index (RS) to measure the degree of interdisciplinarity for each institute (aggregate over all publications of the institute) and experiment with different dissimilarity measures. The Rao Stirling diversity is defined as

$$RS(d) = \sum_{i=1}^{K} \sum_{j=1\ (i \neq j)}^{K} P(i|d)\ P(j|d)\ \ \delta(i,\ j)$$

In addition, the broadness of an institute can be determined by means of the Shannon Entropy (H) based on the distribution over latent topics for each institute. The measure combines the variety and balance dimension, while it ignores disparity. A high topic entropy signals an even distribution and broader spectrum of topics. Shannon Entropy is defined as

$$H(d) = -\sum_{i=1}^{K}\ P(i|d)\ \ln P(i|d)$$

The diversity measure can thus be obtained from the topic-document and word-topic distributions of LDA. More concretely, we use Θ (Topic-Document Probability Matrix) for calculating the balance between topics and Φ (Word-Topic Probability Matrix) for computing the distance δ between topics. A limitation in our use case is obviously, that the underlying distributions are unknown and varying over the parameter setting for the number of topics K and hyper-parameters α and β might yield different Rao scores. Also, the size and length of the training data is crucial, since the priors are estimated from the observed counts in the data.

Datasets

In the present work, we use title and abstract from Scopus, a bibliographic database introduced in 2004 by Elsevier. Scopus provides a comprehensive collection of the scientific landscape, covering the world's leading journals, and is a real-time monitor corpus that is both big in size and rich in metadata. It offers, e.g., research institutions of the authors as metadata records.

Scopus World 2018 (Scopus World)

To explore the interdisciplinarity of an institution, we aim to compute the diversity indicator on a balanced corpus that covers all scientific fields. Therefore, we sampled a corpus from Scopus where we seek to give equal weight to all scientific domains to mitigate the minority class problem, since the distribution of papers and journals over disciplines is heavily skewed (e.g., the humanities are underrepresented in the corpus). The result is a corpus of randomly selected publication abstracts and titles from all major fields of Scopus of the year 2018 (see Fig. 1).

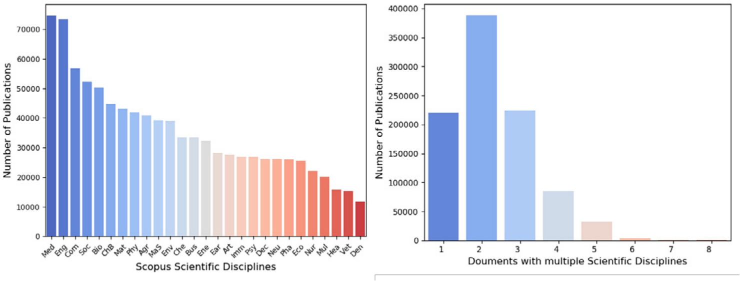


Figure 1: Statistics for Scopus Publications – *Scopus World*

In bibliometrics, the average number of subject categories of a publication, accumulated over an institute, can already serve as an indicator for interdisciplinarity (Levitt und Thelwall, 2008).

The higher the value, the more interdisciplinary the institute. On the publication level, we see that the majority of documents is assigned to more than one discipline, i.e. on average there are 2.3 subject fields per publication (see Figure 1, right).

Institute-specific Publications: Scopus FH and Scopus MPG

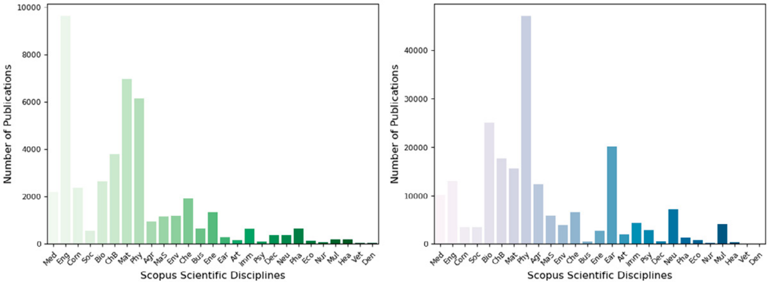


Figure 2. Statistics for *Scopus FH (left) Scopus MPG (right)*

As can be seen in Figure 2, the research profiles of FH and MPG are rather imbalanced, e.g., Scopus FH contains a huge share of publication abstracts from Engineering, while Scopus MPG publishes mostly on Physics and Astronomy. Only a small fraction of articles is dedicated to, e.g., Dentistry. In the FH corpus, 82.41% are assigned to more than 1 field and on average there are 2.47 subject fields per publication, while for the MPG corpus, 70.59% are assigned to more than 1 field and on average there are 2,19 subject fields per publication. Table 2 provides a detailed breakdown of the datasets used in our study.

Table 2. Dataset Statistics

Data Sets	*Number of Institutions*	*Number of Abstracts*
Scopus FH 2010-2018 (Scopus FH)	74	19,661
Scopus MPG 2010-2018 (Scopus MPG)	95	111,986
Scopus World 2018 (Scopus World)		517.516

Empirical Study

Our goal is to test the effects of varying the LDA settings on the diversity measure, composed of disparity, balance and variety. Our research hypothesis is that to provide a good Rao Index of the data, it is desirable that the selected topics are both coherent and interpretable, and have a high coverage of the data.

Choice of the Training and Test Corpora

As training corpus, we use *Scopus World*, and alternatively, *Scopus FH* and *Scopus MPG.* The last two corpora are composed of abstracts from FH and MPG published between 2008- 2018 where we concatenate all abstracts by the same institute to obtain longer documents (with more co-occurrences) that yield better quality topics (Jónsson & Stolee, 2015). We use the institute-specific corpora *Scopus FH* and *Scopus MPG* for testing.

Model Selection and Parameter Settings

Variational inference (Hoffman et al., 2013) as implemented in *gensim* is used for model inference and standard Laplace smoothing factors with $\gamma = 0.1$ and 2,000 iterations. We set the number of topics $K = 100, 150, 200, 250, 300$ topics. As standard parameters of the Dirichlet prior we use a) α = 0.1, b) a non-uniform α estimated automatically from the data (Li et al., 2006) and c) a fixed normalized asymmetric prior of *1/K* (Wallach et al., 2009). Regarding the topics-word distributions, we set a) β = 0.1 and b) β = 0.01, unless otherwise specified.

For pre-processing, we used sentence splitting, tokenisation, lemmatization, and PoS tagging to filter all content words using the Stanford tools[v], keeping only nouns, adjectives, verbs, and foreign words that consist of alphanumeric characters. This resulted in 131,954, 12,598 and 36,381 unique words for *Scopus World*, *Scopus FH* and *Scopus MPG*, respectively.

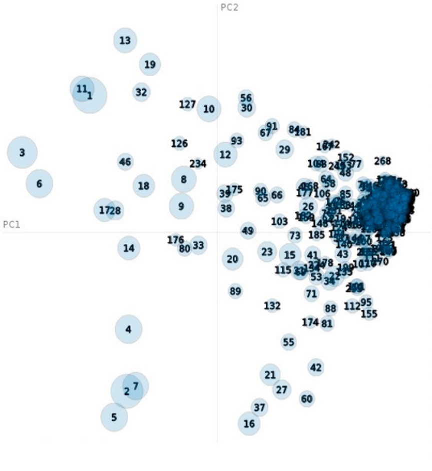
Model Accuracy for Different Settings

We assess modelling accuracy in terms of topic coherence under various settings of hyper-parameters and number of topics. Even though determining the parameters is an established research area and various heuristics exist for real-life applications (Wallach et al., 2009; Lau et al., 2014), Chuang et al. (2012) have shown that a small change in term smoothing and prior selection can significantly alter the ratio of resolved and fused topics. Increasing the number of latent topics often leads to more junk and fused topics with a corresponding reduction in resolved topics. Since LDA results are often difficult to interpret (Chang et al., 2009), we first investigate the outcome of the topic models by humans.

Topic Evaluation by Humans

A qualitative analysis of the topics by manual inspection reveals that there are topics that correspond to scientific domains and others to specific modes of discourse (e.g., description of experimental settings). Topics related to scholarly discourse - signaled by keywords such as *method*, *apply*, or *examine* - are most dominant in the corpus, relative to other topics. In these topics, authors make claims about the key contributions of their paper (Motta et al., 2000). They are more likely in the corpus, because they are relevant for all scientific disciplines.

We also wanted to see how well the topics correlate with an existing categorization schema. To this aim, we compared the topics induced by different LDA models to investigate in how much they correspond to the scientific fields in Scopus ASJC. In order to understand the distribution of topics in terms of size and their overall significance in the corpus, we use the visualization created by LDAvis (Sievert et al., 2014). It also allows assessing how well topics are separated from each other, where topics distance is based on KL divergence.



Alignment	ASJC Category	Topic Words
20,27	Engineering	temperature heat thermal pump
14,25	Materials Science	nanoparticle spectroscopy scanning coating
44	Physics & Astronomy	simulation magnet electron laser controller
12	Chemistry	ratio compound chemical derivative
47	Biochemistry	organic molecular biological fluorescence wavelength
50,193	Computer Science	application tool processing architecture automate
10,18	Medicine	patient cancer therapy tumor survival clinical
9	Chemical Engineering	reaction film synthesis catalyst conductivity
228	Energy	solar atomic newton force graft
20,148	Environmental Science	environmental wind pollution climate atmospheric
76,83	Mathematics	solution equation linear nonlinear
90	Agriculture & Biology	grow leaf fruit crop cultivar farming
145	Pharmacology	medication prescription risk pharmacy
261	Business	investment financing portfolia profitability
10	Immunology & Microbiology	tumor chemotherapy cancer patient
53,87	Social Sciences	survey questionnaire respondent
248	Decision Sciences	decision making rural household
24	Neuroscience	syndrome fatigue disease dysfunction
39	Earth & Planetary Sciences	natural seismic earthquake hazard typology landfill
26,100	Health Professions	model cell method patient
291	Arts & Humanities	artistic artist foam passive hardness
12,46	Economics	real firm sector economy bank commercial
56,187	Psychology	disorder mental disturbance psychiatric psychological
0	Nursing	patient treament analysis time test
165	Veterinary	veterinary life living mikcrobiota
25	Dentistery	implant ceramic arch crown screw dental

Figure 3. LDA Intertopic Distance Map

Figure 3 depicts how LDA topics can be aligned to ASJC codes as reference scientific domains in an overlay representation based on human classification. The topics are drawn from a relatively large *Scopus World* model that was able to uncover a high percentage of scientific topics, covering all ASJC topics (i.e., setting $K = 300$, α asymmetric, $\beta = 0.01$).
Opposed to this, models trained on *Scopus FH* or *Scopus MPG* yielded many uninterpretable and fused topics with a low coverage of ASJC topics.

Topic Coherence versus Coverage

The semantic coherence of the topics is measured using word co-occurrences within the original corpus by the *UMass* coherence score on the top 15 words from each topic (Mimno et al., 2011; Röder et al., 2015). We compare the coherence scores for varying model size of LDA trained on *Scopus World, Scopus FH* and *Scopus MPG.* The LDA models trained on *Scopus World* reach an average UMass score between -7.53 ($K = 100$) to -11.74 ($K = 300$) that decreases as we learn more topics. Even though LDA models trained on *Scopus FH* and *Scopus MPG*, and thus less data, achieve higher UMass scores, they are inferior to the *Scopus World* LDA model in terms of coverage (see Table 3).

Table 3. Coherence and Coverage for varying model size of LDA.

Num Topics	*100*	*150*	*200*	*250*	*300*
Scopus Worlds - Average C_{UMass}	-7.53	-8.77	-9.75	-10.90	-11.74
Scopus FH - Average C_{UMass}	-0.83	-0.91	-0.95	-0.95	-0.98
Scopus MPG - Average C_{UMass}	-3.69	-3.41	-3.29	-3.26	-3.01

Experiments to assess the different dimensions of the Rao Diversity Index

Variety & Balance Scenario: We computed the evenness of the document-topic distribution for all FH institutes under various settings using Shannon Entropy, i.e. a high entropy signals *interdisciplinarity*.
Our experiments on *Scopus FH* used for training and testing show that Shannon Entropy ranges between 0.001 and 1 when averaged over all topics. Results confirm that the choice of α impacts on the entropy values: Setting α to a value closer to zero results in a non-uniform document-topic distribution and lower entropy. Likewise, setting α = *asym* instead of α = *auto* confirms a second hypothesis: the first setting has the effect that the probability mass of the distribution will concentrate on fewer topics per document: Accordingly, entropy values are constantly lower for all topics (see Table 4). Additional experiments demonstrate the impact of proper sampling: Institutes show much higher equality and tendency to focus on more topics when the LDA model is computed on a data set, where samples were drawn such as to accommodate for balance beforehand, i.e. *Scopus World.* Table 5 shows that this results in high entropy values of 0.908 (when averaged over all topic settings). More crucially, however, is the fact that in all cases Spearman's correlation is weak, and Pearson indicates only moderate correlation.

Table 4. Mean Shannon Entropy and Spearman/Pearson - Setting *α=auto/asym* and *α =0.1/ 0.01*

K	*100*	*150*	*200*	*250*	*300*	*K*	*100*	*150*	*200*	*250*	*300*
α=0.1	0.127	0.118	0.114	0.118	0.098	**S.**	0.189	0.146	0.209	0.261	0.285
α=0.01	0.072	0.072	0.072	0.072	0.072	**P.**	0.587	0.638	0.667	0.642	0.724
α=auto	0.099	0.095	0.077	0.058	0.075	**S.**	0.030	0.276	0.327	0.228	0.219
α=asym	0.113	0.067	0.067	0.067	0.067	**P.**	0.393	0.688	0.712	0.738	0.662

Table 5. Mean Shannon Entropy and Spearman/Pearson – for two LDA models (*α = auto)*

K	*100*	*150*	*200*	*250*	*300*	*K*	*100*	*150*	*200*	*250*	*300*
Scopus FH	0.099	0.095	0.077	0.058	0.075	**S.**	0.071	0.243	0.28	0.194	0.238
Scopus World	0.918	0.913	0.919	0.911	0.907	**P.**	-0.02	-0.13	0.101	0.087	0.127

Disparity Scenario: We observed that topical distance decreases, when β approaches 0. The inferred topics are a mixture of multiple topics and less separable when β=0.1 instead of β=0.01. Figure 5 shows the degree of semantic similarity between the topics' word distributions for 100 topics and varying β, using Jensen-Shannon as the distance measure. Pairwise dissimilarity of topics is equally high for all other investigated distance metrics, i.e. Jaccard, Hellinger, Cosine. For a model setting with 100 topics we receive Jensen-Shannon scores of 0.98 on average, ranging between 0.898 and 1 for β =0.01 versus 0.79 and 1 for β=0.1, respectively. Furthermore, the data is more separated when the number of topics becomes larger for both β =0.1 / 0.01.

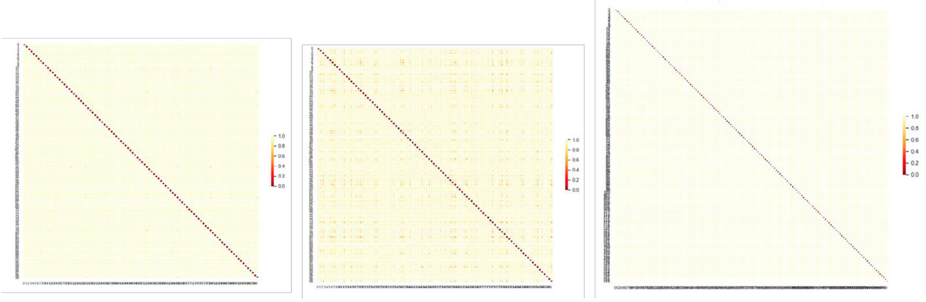


Figure 5. Pairwise Topical Distance based on JS. Topics become dissimilar when β approaches zero and the number of topics becomes larger (*K*=100, β=0.01; *K*=100, β=0.1; *K*=150, β=0.01) (left to right).

Rao Scenario: We investigated the impact of different topic models on the Rao index. First, we calculated the index on the output of the LDA models trained on the institute-specific corpora *Scopus FH* and *Scopus MPG*. In the experiments, we could observe sharp fluctuations of the Rao index when varying over the number of topics (see Fig. 6, left).

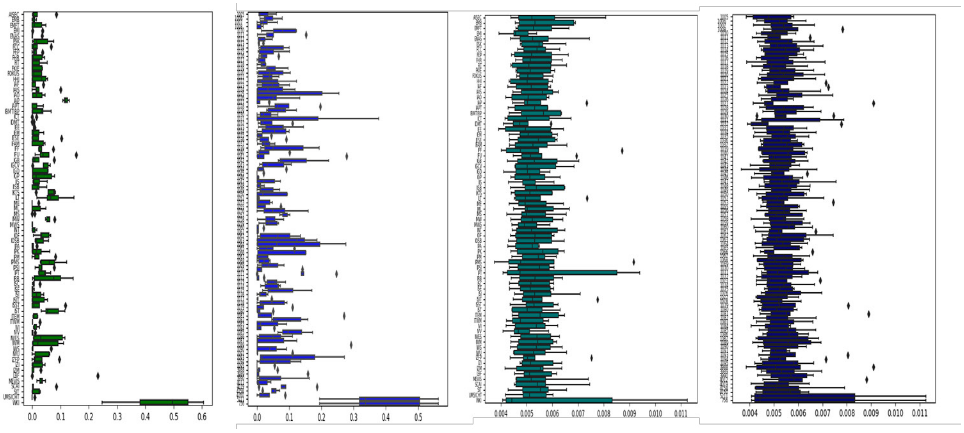


Figure 6. Rao-Index for all Fraunhofer (green) and MPG (blue) institutes; Rao Index is computed for 100, 200, 300 topics on different LDA outputs, i.e. models are trained on *Scopus FH* versus *Scopus MPG* (left) vs. Rao Index computed on *Scopus World* (right)

Rao index values range between 0.001 to 0.605 and 0 to 0.562, with a standard deviation of 0.062 and 0.077 for FH and MPG, respectively. Note that in this case, it was not possible to map the LDA topics fully to all ASJC fields, since the models have a relatively low coverage.
We also calculated the index for various LDA models trained on the *Scopus World* and applied it to the FH and MPG corpora. The setting also makes comparisons between institutes possible and the LDA classifier is less prone to overfitting. However, as shown in section 4, topic quality in terms of qualitative (human judgments) and quantitative (coherence) evaluation showed that many topics were not interpretable or meaningful.
For this setting, the standard deviations are much smaller. In this case, the Rao index takes small values, ranging between 0.004 and 0.011 for both institutes, and thus there is little difference between the values (see Fig. 6, right). The text-based Rao index thus suffers from the same limitations of low discriminating power as the bibliometric-based approach.

Last but not least, we calculated the Spearman and Pearson Rank Correlation of the Rao Index varying on the number of topic and model size. Figure 7 shows the visualization of the coefficients based on the various outputs of Rao, depicting the pairwise correlations as a heatmap. As can be seen, the choice of *K* has a great influence on the Rao results: Pairwise comparisons of Rao results vary a lot, showing that there seems to be no association between the variables. In particular, Spearman correlation is weak, showing that the general rankings amongst institutes is not preserved when varying on the number of topics.

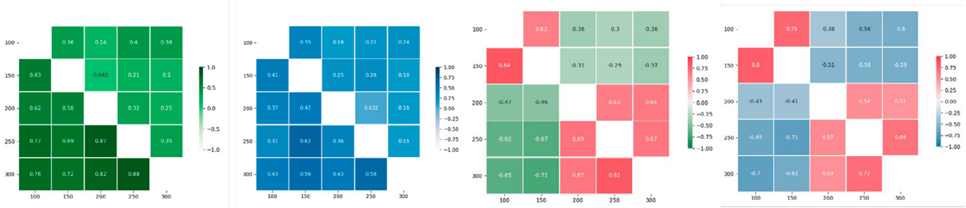


Figure 7: Spearman (upper) and Pearson (lower) Correlation of Rao Index
Computed from various LDA outputs, varying on the number of topics (on x-axis, y-axis) and model size of LDA (small models: left, large models: right), i.e. models are trained on *Scopus FH (green)* versus *Scopus MPG* (blue) vs. Rao Index computed on *Scopus World* (tested on *Scopus FH (*green-red), *Scopus MPG (*green-blue))

Conclusion

In this paper we investigated the Rao indicator for interdisciplinarity based on LDA for two German research institutes. Both institutions are specialised in certain scientific fields and have a more or less high propensity towards interdisciplinary research. It would be a benefit for politicians and decision makers to have an indicator that is able to truly reflecting this trend and which can be computed automatically from any data set.
Yet, our experiments show that the LDA-based Rao metrics has serious limitations and due to its instability might not be a useful indicator of interdisciplinary. Contrary to Bache (2013), who claim that the method could be applied fully automatically and would be largely invariant to the number of topics in the model, our experiments on Scopus and two major German research associations show the opposite. It results in sharp fluctuations that make it an unreliable indicator. We could not find a strong correlation between Rao results that have been generated from different settings. In fact, all parameter variations seem to have a strong effect on the output, i.e. choice of the number of topics, hyper-parameters, and size and balance of the underlying data used for training the model.
There seems to be a consensus in the research community that in order to select the best value of *K*, a qualitative evaluation of the performance of alternative LDA models with varying *K* is required (Suominen, 2016), ensuring that the topic model is able to represent and cover all major scientific fields. Moreover, it is crucial that hyper-parameters are set in such a way that

they produce a topic model with sparse topic and word distributions. A qualitative analysis of the topics of various models reveals that the models fail to differentiate scientific topics from scientific discourse and junk topics. However, topics related to scholarly discourse not necessarily indicate interdisciplinary studies (apart from Scientometrics).

References

Aletras, N., & Stevenson, M. (2014). Measuring the similarity between automatically generated topics. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers* (pp. 22-27).

Bache, K., Newman, D., & Smyth, P. (2013). Text-based measures of document diversity. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 23-31).

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE signal processing magazine*, *27*(6), 55-65.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Brzezinski, M. (2015). Power laws in citation distributions: evidence from Scopus. *Scientometrics*, *103*(1), 213-228.

Cassi, L., Champeimont, R., Mescheba, W., & De Turckheim, E. (2017). Analysing institutions interdisciplinarity by extensive use of Rao-Stirling diversity index. *PloS one*, *12*(1), e0170296.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, *22*, 288-296.

Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 443-452).

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22-29.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235.

Guo, Z., Zhu, S., Chi, Y., Zhang, Z., & Gong, Y. (2009). A latent topic model for linked documents. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 720-721).

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 363-371).

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, *14*(1), 1303-1347.

Jagarlamudi, J., Daumé III, H., & Udupa, R. (2012). Incorporating lexical priors into topic models. In *Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 204-213).

Jónsson, E., & Stolee, J. (2015). An evaluation of topic modelling techniques for twitter.

Lau, J. H., Newman, D., & Baldwin, T. (2014, April). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 530-539).

Levitt, J. M., & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. *Journal of the American Society for Information Science and Technology*, *59*(12), 1973-1984.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348-362.

Leydesdorff, L. (2018). Diversity and interdisciplinarity: how can one distinguish and recombine disparity, variety, and balance?. *Scientometrics*, *116*(3), 2113-2121.

Leydesdorff, L., Caroline S. Wagner C., Bornmann, L. (2019) Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient, *Informetrics,* 13 (1).

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning* (pp. 577-584).

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262-272).

Motta, E., Shum, S. B., & Domingue, J. (2000). Ontology-driven document enrichment: principles, tools and applications. *International Journal of Human-Computer Studies*, *52*(6), 1071-1109.

Nanni, F., Dietz, L., Faralli, S., Glavaš, G., & Ponzetto, S. P. (2016). Capturing interdisciplinarity in academic abstracts. *D-lib magazine*, *22*(9/10).

Newman, D., Bonilla, E. V., & Buntine, W. (2011). Improving topic coherence with regularized topic models. *Advances in neural information processing systems*, *24*, 496-504.

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, *100*(3), 741-754.

Paul, M., & Girju, R. (2009). Topic modeling of research fields: An interdisciplinary perspective. In *Proceedings of the International Conference RANLP-2009* (pp. 337-342).

Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. In *Twenty-fourth international joint conference on artificial intelligence*.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, *82*(2), 263-287.

Ramage, D., Manning, C. D., & McFarland, D. A. (2010). Which universities lead and lag? Toward university rankings based on scholarly output. In *Proc. of NIPS Workshop on Computational Social Science and the Wisdom of the Crowds*.

Rao, C. R. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, *21*(1), 24-43.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408).

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, *4*(15), 707-719.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(10), 2464-2476.

Syed, S., & Spruit, M. (2018, April). Selecting priors for latent Dirichlet allocation. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)* (pp. 194-202). IEEE.

Talley, E. M., Newman, D., Mimno, D., Herr II, B. W., Wallach, H. M., Burns, G. A., ... & McCallum, A. (2011). Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, *8*(6), 443.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014, January). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International Conference on Machine Learning* (pp. 190-198).

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., ... & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of informetrics*, *5*(1), 14-26.

Wallach, H., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in neural information processing systems*, *22*, 1973-1981.

Wang, K., Sha, C., Wang, X., & Zhou, A. (2014). Based on citation diversity to explore influential papers for interdisciplinarity. In *Asia-Pacific Web Conference* (pp. 343-354). Springer, Cham.

Wang, Q., & Schneider, J. W. (2020). Consistency and validity of interdisciplinarity measures. *Quantitative Science Studies*, *1*(1), 239-263.

Wang, X., Fang, A., Ounis, I., & Macdonald, C. (2019). Evaluating Similarity Metrics for Latent Twitter Topics. In *European Conference on Information Retrieval* (pp. 787-794). Springer, Cham.

Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767-786.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology*, *67*(5), 1257-1265.

Zhao, H., Du, L., Buntine, W., & Liu, G. (2017). MetaLDA: A topic model that efficiently incorporates meta information. In *2017 IEEE International Conference on Data Mining (ICDM)* (pp. 635-644). IEEE.

Zhou, Q., Rousseau, R., Yang, L., Yue, T., & Yang, G. (2012). A general framework for describing diversity within systems and similarity between systems with applications in informetrics. *Scientometrics*, *93*(3), 787-812.

[i] https://www.nsf.gov/od/oia/additional_resources/interdisciplinary_research/definition.jsp

[ii] In scopus, e.g., the classification of scientific papers is derived from the classification of journals and considers 27 top-level ASJC codes or 334 sub-level ASJC codes

[iii] Brezezinski (2015) find evidence for power laws in the citation distributions from Scopus.

[iv] The distance between categories can be calculated by the means of a matrix of citation flows between categories (Rafols and Meyer, 2010). A common measure is cosine similarity.

[v] https://stanfordnlp.github.io/CoreNLP/

Gendered feedback and negative adjective use in peer-reviewer reports

Alessandra Zimmerman[1], Helen Greaves[2], Richard Klavans[1], Jonathan Best[3] and Gemma E. Derrick[2]

[1] *azimmerman@proposalanalytics.org and rklavans@proposalanalytics.org*
Proposal Analytics Inc, 606 Wayne St, Wayne, PA, United States

[2] *h.greaves@lancaster.ac.uk and g.derrick@lancaster.ac.uk*
Centre for Higher Education Research and Evaluation, Lancaster University, Lancaster, United Kingdom

[3] *j.best@wellcome.org*
The Wellcome Trust, Euston Road, London, United Kingdom

Abstract

This research in progress reports on an ongoing analysis of submissions and review reports made to Early Career Researcher (ECR) fellowship applications made to the Wellcome Trust between 2009-2019. It combines a small scale, linguistic coding of reviewer reports; interviews with unsuccessful candidates; and a large-scale, text-driven analysis of noun phrases surrounding the use of gendered pronouns (he/his-she/her-they/their/you) within peer review reports provided to applicants as feedback after receiving the outcome. This research in progress reports on the initial findings of the research garnered from the small-scale linguistic coding of review reports, and interviews with unsuccessful applicants. The initial analysis shows that, on average, women receive longer reports than men. Additionally, candidates identified as women by the reviewer's commentary on average receive 16.18 negative comments compared to 9 for those identified as men. In addition, many negative comments for women are based on assessments of personal competence and measures of performance up until the time of the application in addition to the project assessment, whereas negative comments for men are focused on characteristics of the proposed research project. When triangulated with one-on-one interviews with male and female candidates who were unsuccessful, where present, gendered notions in reviewers report provided as feedback were identified by applicant and this influenced the applicants' future career decisions, as well as how the applicant perceived the fairness of the evaluation governed by Wellcome, as well as the use of peer review in academia more generally. The large-scale analysis of noun phrases is ongoing, will feed into the analysis briefly discussed above, and will form part of this presentation in July.

Introduction

The peer review process used to aid the decision of competitive grant funding is a cornerstone of academic governance. When functioning well, it provides funding agencies with tools to aid decision-making that is driven by expert and/or peer evaluations, and it provides applicants with feedback and performance assessment that can ideally be used to strengthen future research activities and decisions regardless of the outcome. However, when functioning badly, the single-blind peer review process adopted by the majority of funding agencies (Tomkins et al, 2017) allows peer reviewers to exercise unconscious bias blindly, with little repercussions for the reviewer for when these biases subsequently influence outcomes. In addition, how applicants receive these 'signals' (Derrick et al, in progress) in science influences how their future research and career choices, as well as decreases their trust in the peer review process and personal sense of social justice.

It is widely recognized that gender bias is widespread in academia and academic institutions, including peer review (Monroe et al, 2008; de Paola & Scoppa, 2015) Addressing this bias, however, is not straight-forward with its unconscious forms influencing how research, and research careers are governed; and expectations of performance is assessed, globally. By examining reviewers reports and using interviews with unsuccessful candidates for ECR fellowship applications at The Wellcome Trust (2009-2019), this research in progress investigates how gender-bias is manifested in how reviewers assess researchers and their research proposals.

Methods

Survey and Wellcome Trust database

A database of proposals submitted to the Wellcome Trust, and their relative reviewer reports for both successful and unsuccessful applicants from 2009-2019 was constructed. Five funding fellowship calls were chosen due to their focus on ECR research, these included;

1. Sir Henry Wellcome Postdoctoral fellowships;
2. Research Fellowships in the Humanities and Social Sciences;
3. Sir Henry Dale Fellowships;
4. Research Career Development Fellowships; and
5. Starter Grants for Clinical Researchers.

From these proposals (n=4109) a survey of respondents (n=411) was then used to query whether a resubmission after an initial (T_0) failure at the Wellcome Trust was attempted; to establish how many resubmissions were made prior to success ($T_1 \ldots T_x$); and to collect subsequent funding applications.

For reasons outlined in (Derrick et al, in progress), this project focused on analysing the review and provision of feedback for unsuccessful applications only. From the survey, unsuccessful applicants who had reapplied; (1) Once after T_0 (=T_1); (2) more than once after T_0 (T_0 + x (T_x)); and (3) had not reapplied, were identified. Feedback provided to each group of applicants was then collected for the linguistic analysis, and respondents invited for an interview with the research team. Publication data on all participants (n=4109) was also collected and used to track career data following the initial application (T_0) to The Wellcome Trust.

Qualitative coding of reviewer comments

From the initial database, and cross referencing with data from the survey, 27 documents from ECR fellowship applications made to the Wellcome Trust, comprising 81 reviewer comments, were extracted and thematically-coded using NVivo11. Coding was linguistically-informed and built upon conversations between the research team and the Wellcome Trust and coded as: (1) negative; (2) positive; or (3) questionable[1].

In parallel, reviewer comments were coded for their use of pronouns associated to the applicant ((1) he/his; (2) her/she; (3) they/their/you)) within sentences. In addition, coding for pronouns also included the entire review if the reviewer announced the gender of the applicant continuously throughout their comments (Male (M) or Female (F)). If the gender is not announced (they/their/you), the comments were coded as Non-Gendered.

Sections of the reviewer comments were also coded where comments discussed the (1) candidate (*Candidate*); (2) research setting (both on an institutional level or in reference to a particular supervisor) *(Setting*); and (3) discussion of the proposed project (*Project*).

Using the compound query function in Nvivo11, the instances of overlap between the location of an emotional response coding and the 3 topic identifiers listed above were determined for each reviewer's comments. The instances of overlap where then separated based on whether or

[1] 'Questionable' was a category where the statement could not be interpreted as 'negative' or positive' but where the response from the applicant to the reviewer's comment would from the applicant was anticipated to be associated with a confused response, or where the negative connotations could have been an accident on the part of the reviewer's word choices.

not the reviewer had explicitly used gendered pronouns in their text. The sums of the instances of overlap were then divided by the number of reviewers for each gender.

The emotive response to reviewer comments was triangulated with interviews with the applicants, and their description of their response to the feedback was feedback to enhance the coding of the reviewer comments.

Findings

The preliminary results showed that reviewers that used female pronouns in reference to the candidate wrote on average 16.19 comments that could potentially evoke a negative emotional response, 15.22 positive, and 1.63 questionable.

In addition, for reviewers that used gendered pronouns in the review, the use of male pronouns was associated with, on average, 9 negative comments, 12.91 positive comments, and 1.26 questionable comments. In comparison, comments that used no gendered pronouns (they/their/you) had 5.5 negative, 3.85 positive, and 0.4 questionable.

Table 1. The number of comments by reviewers that could elicit potential emotional responses for applicants in the sections that discuss the applicant, their research setting, and the proposed project – analyzed by identified gender.

Table	*Candidate*			*Setting*			*Project*		
	Neg	*Pos*	*Off*	*Neg*	*Pos*	*Off*	*Neg*	*Pos*	*Off*
Female	1.85	3.74	0.41	0.11	2.74	0.41	11.85	6.40	0.78
Male	0.56	3.88	0.29	0.38	2.24	0.09	7.6	5.53	1.01
Non Gendered	0.45	1.1	0.05	0.5	0.6	0.05	3.5	1.6	0.25

The preliminary linguistic analysis for negative emotional words of the larger dataset by applicant-identified gender, showed a small trend towards the use of more negative adjectives in situations where the applicant's gender was specifically referred to by the reviewer (e.g. he/his; she/her). It should be noted that the occasional negative comment is excepted, if not beneiuficial to the applicant's development. It is in the frequency and intensity of the negative comments that issues may arise.

> *"The 'idea' of a flowchart and checklists, though, has already been pioneered in patient safety in surgery by a* medical professional." (emphasis theirs) (2.36% of reviewer comments were a direct attack, 19.15% were generally negative)[2]
>
> *"First of all, the applicant does not demonstrate that she has any knowledge of the social science literature on areas such as…"* (3.79% of reviewer comments were a direct attack, 28.73% were generally negative)
>
> **Reviewer Feedback for female applicants**
>
> ***"****However, the proposal as currently constituted appears premature and insufficiently though through. "(*0.35% of reviewer comments were a direct attack, 11.29% were generally negative)
>
> **Reviewer Feedback for a male applicant**

Interviews with applicants exploring their response to the language of reviewer reports, found that in many cases, the negative framing of the comments elicited negative reactions from

[2] Percentages are derived by NVivo as percent of total characters in the document coded to that category. This includes the form email prior to the review comments, as well as all reviewer comments submitted to the applicant.

applicants. This was particularly the case for female applicants, who are reportedly sensitive to feedback identified as 'gendered' by the above thematic coding, regardless of whether the focus of the feedback was on either the Candidate, Setting, or Project. For example, for a Female candidate;

> *However, the application, as it now stands, does not do enough to highlight her career progression or connect the dots between her past, present and future career aspirations. Put simply, it does not explain where she has been and why and how this links with where she is going. There is a complicated mix of broadcasting, journalism, publishing and law training. It appears erratic without explanation and discussion (the form provides several places to do this which are not used well). There is a problem with structure (and good use of word limitations) to convey the applicant's suitability to a fellowship and how this fits with her past and future academic career intentions.*
>
> **Reviewer feedback**

Compared to a Male candidate on the same topic;

> *His track record shows promise due to a laudable willingness to learn and employ cutting edge techniques and use different experimental organisms to address the topic being addressed.*
>
> **Reviewer feedback**

And;

> *His dual training in medicine and law, along with his experience and high-level training as a physician, will no doubt serve him well throughout his career.*
>
> **Reviewer feedback**

In some cases, it was difficult to ascertain whether the review had simply overlooked polite convention related to the use of salutations;

> *"It is terrific that Ms. **** is able to produce translations of Persian works that will no doubt of interest to future scholars and be significant to her own work."*
>
> **Reviewer feedback** - (every other reviewer called her Dr.)

The negative comments are not the only oness with a gendered imbalance – the effusiveness of praise can also vary based on the pronouns used in the review.

> *"From my exerience I would consider this at the international forefront of work in this area and am not aware of better modelling work of this nature ongoing elsewhere ."*
>
> **Reviewer Feedback for a male applicant**

> *"This is strong proposal that by collaboration takes an interdisciplinary approach to an important research question. I think the methods are entirely feasible."*
>
> **Reviewer Feedback for a female applicant**

Or whether this was examples of the operationalization of embedded gender bias in research culture more generally. The large-scale, linguistic word analysis will be used to explore this further.

Discussion

The input of external expert (peer) assessments is a hallmark of the peer review process. At the Wellcome Trust, external review reports are considered by a separate panel assessment process that uses these reports to aid their decision-making about who will successfully receive research funding, and those that will be unsuccessful.

Although, the results cannot indicate if the funding outcomes were influenced by gender bias, the analysis of the way that gendered language in reviewer comments on unsuccessful grant applications is associated with positive and negative comments about the applicant, their environment and research project, is concerning. In addition, it cannot reliably be concluded that a higher proportion of negative comments used in comments is associated with either negative outcomes for the applicant, or the existence of gender-bias in the decision-making process of peer review panels; it is known that the authority that these peer review reports hold over these decision-making processes is large. Therefore, these results show the potential of how unconscious gender-bias of reviewers can influence the language used in peer review reports and by supplying these reports to the decision-making panel could be used to influence outcomes. This is especially the case if the reviewer reports are received by the panel as objective assessments of research value and therefore seemingly free from influences of bias. In this way, the influence of gender-bias in academic decision making is not just unconscious, but also embedded into systems of academic governance and evaluative practice.

Acknowledgments

This project was supported by funding from the Wellcome Trust as part of the project "Minimising the career cost of first failure: Creating a peer review process that treats failure as a learning tool for ECRs".

References

De Paola, M., Scoppa, V. (2015) Gender discrimination and evaluators' gender: evidence from Italian academia. *Economica* 82.325 (2015): 162-188. https://onlinelibrary.wiley.com/doi/full/10.1111/ecca.12107

Derrick, G.E. Greaves, H., Zimmerman, A., Klavans, R., Best, J. (in progress) Signals in science matter: Categorising the effect of negative review reports on applicants' future career decisions.

Monroe, K. et al. (2008) Gender Equality in Academia: Bad News from the Trenches, and Some Possible Solutions. *Perspectives on Politics*, vol. 6, no. 2, 2008, pp. 215-233. *JSTOR*, www.jstor.org/stable/20446691.

Tomkins, A., Zhang, M., Heavlin, W.D. (2017) Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences* Nov 2017, 114 (48) 12708-12713; DOI: 10.1073/pnas.1707323114 https://www.pnas.org/content/114/48/12708

A conceptual evaluation of measures of interdisciplinarity

Sander Zwanenburg[1], Maryam Nakhoda[2] and Peter Whigham[3]

[1] *sander.zwanenburg@otago.ac.nz*
Department of Information Science, University of Otago, New Zealand

[2] *maryam.nakhoda@postgrad.otago.ac.nz*
Department of Information Science, University of Otago, New Zealand

[3] *peter.whigham@otago.ac.nz*
Department of Information Science, University of Otago, New Zealand

Abstract

While interdisciplinary research has attracted much attention, this has not yet resulted in a coherent body of knowledge of interdisciplinarity. One of the impediments is a lack of consensus on its conceptualization and measurement. Some of the proposed measures have shown to misalign empirically, meaning that conclusions about IDR can differ across measures. To clarify this disagreement conceptually, and to stimulate better coherence in measurement, this paper starts with a review of the IDR definitions. From a synthesis of these definitions, we provide a conceptual definition of the construct along with evaluation criteria for its measures. We use these to evaluate current measures of IDR. The results show that most measures meet the usability criteria, but measures vary widely in meeting other criteria, which can explain some of the observed inconsistencies in earlier studies. The "Integration score" from Rafols and Meyer performed best. By using the results in selecting measures, researchers can, in our view, draw more consistent conclusions, aiding in the development of a coherent body of knowledge of this ever-important phenomenon.

Introduction

Many complex problems arising in society require the integration of knowledge beyond the boundaries of a single discipline (Wang et al., 2017; Choi & Richards, 2017; Okamura, 2019, MacLeod & Nagatsu, 2016), such as climate change, food security, peace, social justice (Al-Suqri et al., 2017), and environmental issues (Steele & Stier, 2000; Morillo, Bordons, & Gómez, 2003). Engaging in interdisciplinary research (IDR), however, is challenging. Specialised disciplines have defined research practices, languages, philosophies and communities that often seem inconsistent or incompatible. Understanding these problems and how they are to be overcome is best achieved by meta-research.

However, to date, studies on IDR have not constructed a coherent body of knowledge due to a lack of consistency in measuring interdisciplinarity (Leydesdorff & Rafols, 2011; Wang & Schneider, 2020). Choosing data sets and methodologies produce "inconsistent and sometimes contradictory" results (Digital Science, 2016, p. 2; Wang & Schneider, 2020). Partly this may be due to different indicators capturing different aspects of interdisciplinarity (Leydesdorff & Rafols, 2011). For example, they may capture one or multiple of the three dimensions of diversity: variety, balance, and disparity (Leydesdorff, Wagner , & Bornmann, , 2019). Some take into account the varying distances between fields; others do not. Some assume a network view, and others a more hierarchical structuralists perspective (Wagner et al., 2011). Yet even when measures purportedly capture similar aspects, they may still be empirically inconsistent (Wang & Schneider, 2020).

These inconsistencies stem from diversity in conceptualization and operationalization of interdisciplinarity. First, some authors (e.g., Karlqvist, 1999) have approached interdisciplinarity qualitatively, others quantitatively. The range of quantitative conceptualization of interdisciplinarity covers two distinct perspectives. One is considering the processes and dynamics that contribute to interdisciplinary research and the other one focuses on research publications and to what extent they integrate different research disciplines (

Mugabushaka, Kyriakou, & Papazoglou,, 2016). Other operationalizations provide more details about the integration of different fields of knowledge, in terms of methods, tools, theories, and data (Rafols & Meyer, 2010; Porter et al., 2007). Each of these approaches can lead to different definitions, measures and indicators. Thus, the choice of a specific view and its operationalized indicators produce inconsistent results (Wang & Schneider, 2020).

If IDR would be quantified with more consistent measures, it is more likely that conclusions about this phenomenon, including how to maximise its success, will also be consistent, allowing for a more coherent body of knowledge.

In this paper we aim to take a first step towards this. We first review and synthesize the conceptual definitions of IDR. Based on this we develop conceptual evaluation criteria and apply them to extant measures. We then discuss the results in the context of empirical evaluations elsewhere, and how researchers can choose the best measure toward more consistency and coherence in developing an understanding of IDR.

Definition of IDR

Our review of 25 definitions of interdisciplinarity provided in the literature[1] reveals various similarities. Within the context of quantitative, each definition of interdisciplinarity either explicitly or implicitly refers to multiple disciplines, or similarly, *fields*, or *bodies of knowledge or research practice*. Regardless of the name given to them, they are treated as distinct collections (or systems) of ideas, knowledge, data, tools, theories, practices, specialists, or combinations thereof. Central to most definitions is a connection, communication, or exchange of such items across distinct collections, and/or a resulting intellectual synthesis, fusion, or integration. This result may serve different purposes, but typically relate to understanding issues or solving problems that are not contained to a single discipline.

Most definitions imply these collections must be distant, distinct, or diverse, either sufficiently so for interdisciplinarity to 'occur' or 'apply', or increasingly so for IDR to be more evident or pronounced. Similarly, some definitions imply that a condition for IDR is that the resulting synthesis or integration must be of a sufficient degree or depth. Most definitions do not carry assumptions or criteria about those conducting IDR, such as teams versus individuals, or specialists versus generalists. They also do not tend to constrain IDR as an attribute of particular classes of objects, such as authors, papers, departments, or fields of study.

Based on this review, we define IDR as *the integration of knowledge from diverse disciplines*. This definition allows for two main conceptual dimensions of ID: the extent to which knowledge is integrated, and the diversity of the source disciplines. Yet the definition itself carries no specific implications for the scales of diversity, integration, or interdisciplinarity itself. Nor does it carry implications for the relationship between interdisciplinarity and its two dimensions. In line with the reviewed definitions, this definition allows for describing various classes of objects. Individual papers, authors, author teams, departments, universities, journals, conferences, fields of study and entire disciplines can all be described in terms of the extent they achieve or reflect the integration of knowledge from diverse disciplines. The components of this definition can be understood as follows:

– *A discipline* is a distinct collection of facts, concepts and methods (Braun & Schubert, 2003; Barry, Born, & Weszkalnys, 2008). Therefore, discipline refers to any area of study that has had some level of coherence, by having defined communities, educational programmes,

[1] We searched for keywords such as "interdisciplinarity" and "interdisciplinary research" within academic databases (e.g., WOS, Scopus, and ProQuest). Around 100 sources including books, journal papers, and research reports were retrieved from which we retrieved 25 definitions from various sources. For space considerations, they are not reproduced here.

research culture, and a shared body of knowledge. Since the boundaries of disciplines can vary across communities and geographies, change over time, and be expressed at different levels of granularity, there is no universally accepted or persistent classification of definitions.
- *Diversity,* in the context of disciplines, refers to the variety, balance, and disparity of a set of disciplines (Leydesdorff & Rafols, 2011; Zhou et al., 2012). These dimensions or aspects of diversity are additive: the higher the number of different disciplines in a set, the more evenly their frequency is distributed, and the more disparate they are from each other, the more diverse is the set. By extension, the more diverse the disciplines from which research integrates knowledge, the more interdisciplinary the research.
- *Integration of knowledge* lies at the heart of the concept of IDR (e.g., Klein, 1990; Rhoten & Pfirman, 2007; Porter & Rafols, 2009; Rafols & Meyer, 2010; Chang & Huang, 2012) and aids in distinguishing IDR from similar concepts such as multi-disciplinarity (Holland, 2008). Based on the review, integration must be broadly understood as any combination, fusion, synthesis, or even juxtaposition of two or more ideas, producing a whole idea, a more general idea, or a meta-idea. While philosophically, it is possible to speak of the depth or degree with which two or more ideas are integrated (Rafols & Meyer, 2010), definitions of interdisciplinarity tend to be more consistent with a dichotomous view: ideas are assumed integrated or they are not. While the integration of knowledge is fundamentally a cognitive process (Wagner et al., 2011), it may emerge both from within individuals and from interactions of multiple individuals (Porter et al., 2007).

Based on this synthesis of the definition of IDR, we aim to evaluate how commonly used measures of IDR align with it.

Method

Our conceptual evaluation of commonly used measures vis-à-vis a synthesized definition of IDR will follow two steps. First, from the preceding review we will draw out conceptual criteria for measures, such that measures that meet these criteria are in alignment with the definition of IDR. Next, we identify commonly used measures, and examine whether they meet these criteria.

Evaluation criteria

Good measures meet a variety of criteria: they can be used where needed (usability), they reflect the construct as a whole (construct validity), they cover the content domain of the construct (content validity), and they can distinguish the construct from similar ones (discriminant validity) (MacKenzie, Podsakoff, & Podsakoff, 2011).

Usability

IDR has been studied in the context of individual papers (Porter et al ,2007; Chen et al. ,2015), authors (Qin, Lancaster, & Allen, 1997), groups of authors such as departments or universities (Bordons et al., 1999; Gowanlock & Gazan, 2013), journals (Leydesdorff, 2007) and disciplines (Rinia, van Leeuwen, & van Raan, 2002; Schummer, 2004). Using a common measure across these studies will help develop a common body of understanding of IDR. Therefore:

1; *Multi-object*: The measure is applicable to a variety of objects of study, including individual papers, authors, institutions, journals, and disciplines.

Most research on IR involves evaluations and comparisons within or across these objects of study (Morillo, Bordons, & Gómez, 2003; Leydesdorff, 2007; Wang & Schneider, 2020). For example, one may wish to compare a new journal to an established journal, or a small department to a larger one. Given that they can vary greatly in the quantity of research they represent, such comparisons are only meaningful when controlling for these quantities.

2; *Size independent*: The measure's values for IDR of objects of study are independent of the amount of research (e.g., number of published papers) belonging to these objects.

Measures of IDR have leveraged various classifications of disciplines, such as the Web of Science Subject Classification (Wagner et al., 2011) and the All Science Journal Classification (Leydesdorff, de Moya-Anegón, & Guerrero-Bote, 2015), or proxies of disciplines, such as journal title. All of these have different granularities. Since the identity of disciplines varies across communities and institutions (e.g., Zhang, Rousseau, & Glänzel, 2015), and changes over time (Huutoniemi et al., 2010), there can be no universally or permanently true classification of disciplines.[2] To accommodate multiple classifications within consistent measurement, measures should be able to be applied to any of these without carrying an inherent bias. In particular, measures should not result in widely different scores of IDR if classifications of different granularities are used.

3; *SC unbiased*: The measure is applicable to any subject categorization and is independent of its granularity.

Construct and content validity

Measures with construct validity capture the meaning of the construct (Zwanenburg & Qureshi, 2019), in this case the degree with which knowledge from diverse disciplines is integrated. This involves two dimensions, namely knowledge integration and diversity in disciplines. Content validity refers to the coverage of the content domain of a construct, namely these two dimensions and their content.

4a; *Integration of sources*: The measure reflects the degree to which knowledge from source disciplines is integrated. [3]

Fully covering knowledge integration requires the identification of all disciplinary sources of knowledge that have been relied on to develop the research. It also requires the degree to which knowledge from each source discipline is integrated.

4b; *All sources identified*: The measure is based on a complete identification of disciplinary sources of knowledge that have led to the research.

4c; *Integration of each source*: The measure is based on the degree to which knowledge from each source discipline is integrated.

The content of the diversity construct, as used in the IDR context, consists of disparity, variety, and balance (Wang, Thijs, & Glänzel, 2015).

5a-c; *Disparity, variety, balance*: The measure is sensitive to differences in the (a) disparity, (b) variety, and (c) balance of source disciplines.

IDR measures

From our literature review we included all measures when they were presented as a measure of IDR. This was regardless of them being deployed, the intent or motivation behind their

[3] This goes beyond *representing* multiple disciplines, as is common in studies on multi-disciplinarity.

proposal, or any particular object of study (Wang, Thijs, & Glänzel, 2015). The measures are listed below.

1. *P_multi*: This measure considers "percentage of journals classified in the category analysed and in any other SciSearch category at the same time" (Morillo, Bordons, & Gómez, 2001, p. 205).
2. *p_outside*: This measure indicates "the percentage of journals in i (a WOS subject category) that have been assigned to more than one research area" (Wang & Schneider, 2020, p.242).
3. *Pro*: The percentage of references outside the category of the focal journal (Morillo, Bordons, & Gómez, 2001; Porter & Chubin, 1985)
4. *D_links*: Diversity of links, indicating "the number of links between different SCs [subject categories, red.] established by journals in each category" (Morillo, Bordons, & Gómez, 2003, p. 1241).
5. *Pratt index*: An index to assess frequency distribution to compare journal and subject concentration in different fields (Pratt, 1977).
6. *Specialization index (Spec)*: "This measures the spread of references that publications in SC$_i$ cited over all other SCs" (Wang & Schneider, 2020, p.243). "For the set of journal articles, it counts the number of publications in each SC, squares the count for each SC; and then sums these. Finally it divides that sum by the square of the sum of all the counts" (Porter et al., 2007, p. 138).
7. *Brillouin index*: "A diversity index that combines the concepts of richness (i.e., the number of observations) and relative abundance (i.e., the distribution of observations among categories)"(Steele & Stier, 2000, p.478).
8. *Gini coefficient*: "A measure of inequality or unevenness in a distribution" (Leydesdorff & Rafols, 2011, p. 90). It considers the distribution of references over SCs for a group of publications (Wang, Thijs, & Glänzel, 2015).
9. *Rao-Stirling (RS) diversity index*: A measure of diversity used for interdisciplinarity, including "the number of disciplines cited (variety), the distribution of citations among disciplines (balance), and, crucially, how similar or dissimilar these categories are (disparity)" (Porter & Rafols, 2009, p. 721). Increase in any of these attributes leads to increase in the diversity of the system (Rafols & Meyer, 2010).
10. *Hill-type true diversity index*: Similar to the RS diversity index, the Hill-type indicator gives more weight to variety (Zhang, Rousseau, & Glänzel, 2015).
11. *Coherence*: "The indicator of field coherence quantifies the degree of knowledge integration at the level of fields (or rather subfields), that is, Subject Categories" (Soós & Kampis, 2012).
12. *Betweenness-centrality (BC)*: A measure proposed to assess interdisciplinarity of journals. Betweenness measures the degree of centrality that a node (representing e.g. an article or journal) is located on the shortest path between two other nodes in a network (Leydesdorff, 2007).
13. *Journal interdisciplinarity*: This measure quantifies the degree of specialization versus interdisciplinarity exhibited by a journal based on the bipartite network between scholars and journals (Carusi & Bianchi, 2020).
14. *Author interdisciplinarity*: This measure extends Rao-Stirling diversity measure, with co-author networks and from the perspective of disparity (Zhang et al., 2020).

Results

The outcome of the evaluation is provided in Table 1.

Table 1. A conceptual evaluation of measures of interdisciplinarity

	Criteria						
Measure	*1: Multi-object*	*2: Size Indep.*	*3: SC unbiased*	*4a: Integration of sources*	*4b: All sources identified*	*4c: Integration of each source*	*5a-c: Disparity, variety, balance**
1. Percentage of multiassigned journals (p_multi)	✓	✓	×	×	×	×	×
2. Percentage of journals outside the area (p_outside)	×	✓	×	×	×	×	×
3. Percentage of references outside the category (pro)	×	✓	×	×	×	×	×
4. Diversity of links (d_links)	×	✓	×	×	×	×	×
5. Pratt index	×	✓	×	×	✓	×	1
6. Specialization index (Spec)	×	✓	×	×	✓	×	1
7. Brillouin diversity index	✓	×	×	×	✓	✓	2
8. Gini coefficient	✓	✓	×	×	✓	✓	2
9. Rao-Stirling (RS) diversity index	✓	✓	✓	✓	✓	✓	✓
10. Hill-type measure	✓	✓	✓	✓	✓	×	✓
11. Coherence measure	×	✓	✓	✓	✓	✓	✓
12. Betweenness-centrality	✓	✓	✓	×	×	×	×
13. Journal interdisciplinarity (Carusi and Bianchi, 2020)	×	✓	✓	×	×	×	✓
14. Author interdisciplinarity (Zhang et al., 2020)	✓	✓	N/A	×	×	×	✓**

* '1' and '2' refer to one or two of three criteria being met.

** The three dimensions are tapped into, but not as applied to the diversity of disciplines but that of authors.

Only one criterion is nearly universally met, namely *2: Size independent*, as most measures include a division by e.g., the number of papers published All other criteria are met by three to nine of the 15 evaluated measures. Criterion *4a: Integration of sources* was met least. The typical reason for this was that the consideration of multiple disciplines was not in the context of disciplinary sources. For example, the author-journal bipartite measure taps into clusters of communities, without directly tapping into integration of knowledge. Those that met this criterion referred to the references of articles, papers, or other research works, thus more clearly being based on disciplinary sources of knowledge integration.

There was considerable variability in tapping into, or being sensitive to, the different dimensions of diversity. While variety (simply the number of disciplines) was typically met, disparity was met only by few measures, as most did not account for or were not sensitive to the distance between or similarity of disciplines. This means that scores on these measures are more sensitive to the selection of the classification of disciplines, since these vary in granularity. Several measures, most notably numbered 1-6 which comprise Wang & Schneider's Group 1 of measures, meet few of the criteria, without strong reference to integration or an independence of a chosen classification of disciplines. Several measures met nearly all, and one measure, the Rao-Stirling diversity index, met all of the criteria.

Discussion and conclusions

The literature has shown measures of ostensibly the same construct, interdisciplinarity, do not align empirically (Wang & Schneider, 2020). Our results take a step in identifying the underlying conceptual reasons for this disagreement. To some extent, our evaluation mirrors the empirical evaluation of Wang & Schneider (2020) in identifying two clusters of measures, with similarities in terms of the respective fulfilment of conceptual criteria.

There are several take-aways for researchers interested in operationalizing interdisciplinarity:

- First, to develop a coherent body of understanding on IDR caution is to be exercised in terms of the selection of measures. Our results suggest that the observed empirical disagreement is not just noise but there are significant conceptual differences underlying these measures. The danger of picking a convenient measure is that its conceptual foundations do not align with much of the extant literature and that the results will not be able to contribute to it.
- Second, to exercise such caution means clarifying the conceptual definition of interdisciplinarity. Our review has synthesized 25 definitions found in the literature; this synthesis seems consistent with most of the meaning attributed to interdisciplinarity in the quantitative literature on this phenomenon.
- Third, whether this definition is relied on, or another one, it is imperative to be aware of and communicate the assumptions that this choice implies. For example, by emphasizing variety over other dimensions of diversity will imply tie scores on the measure more strongly to a particular classification of disciplines.

This study provides a handle for researchers wishing more conceptual guidance in the concept and measurement of IDR through the synthesized definition, the identification of evaluation criteria, and the application of these two measures of IDR. While these criteria are not exhaustive in the selection of measures to employ – one can think of practical considerations like the ease with which meta-data is obtained – they do cover the conceptual domain of the construct, and are able to point to important differences in existing measures.

Based on the findings, measures that satisfy the majority of criteria include: Rao-Stirling (RS) diversity index, Hill-type measure, Coherence measure, Betweenness-centrality, and Journal interdisciplinarity (Carusi and Bianchi, 2020). Their common properties are that they use cosine similarity between subject categories, and also include three aspects of diversity (variety, balance, and disparity).

This study does not advocate for any one measure in particular, or show that any of the measures are invalid in themselves. It shows that many do not meet most criteria that are based on a synthesis of the literature. This synthesis is not a consensus on the definition of IDR per se: one can argue for different 'flavours' of IDR, and place more or less importance on its dimensions of integration of knowledge, and the dimensions of diversity of disciplines. However, given the empirical disagreement, a continued emphasis on idiosyncratic definitions, conceptually or operationally, will deter the development of a common body of understanding on IDR in which results and conclusions are intercompatible.

We hope that our synthesis and evaluation help in fostering more consistency in measurement and thus coherence in our body of understanding, such that practitioners and academics alike can unlock more of the great potential of interdisciplinary research.

References

Al-Suqri, M. N., Al-Kindi, A. K., AlKindi, S. S., & Saleem, N. E. (Eds.). (2017). *Promoting Interdisciplinarity in Knowledge Generation and Problem Solving*. IGI Global.

Barry, A., Born, G., & Weszkalnys, G. (2008). Logics of interdisciplinarity. *Economy and society*, 37(1), 20-49.

Bordons, M., Zulueta, M., Romero, F., & Barrigón, S. (1999). Measuring interdisciplinary collaboration within a university: The effects of the multidisciplinary research programme. *Scientometrics*, 46(3), 383-398.

Braun, T., & Schubert, A. (2003). A quantitative view on the coming of age of interdisciplinarity in the sciences. *Scientometrics*, 58(1), 183-189.

Carusi, C., & Bianchi, G. (2020). A look at interdisciplinarity using bipartite scholar/journal networks. *Scientometrics*, *122*(2), 867–894. https://doi.org/10.1007/s11192-019-03309-3

Chang, Y. W., & Huang, M. H. (2012). A study of the evolution of interdisciplinarity in library and information science: Using three bibliometric methods. *Journal of the American Society for Information Science and Technology*, 63(1), 22-33.

Chen, S., Arsenault, C., Gingras, Y., & Larivière, V. (2015). Exploring the interdisciplinary evolution of a discipline: the case of Biochemistry and Molecular Biology. *Scientometrics*, *102*(2), 1307–1323. https://doi.org/10.1007/s11192-014-1457-6

Choi, S., & Richards, K. (2017). Interdisciplinary discourse: Communicating across disciplines. In *Interdisciplinary Discourse: Communicating Across Disciplines* (Issue 2008). https://doi.org/10.1057/978-1-137-47040-9

Digital Science. (2016). Interdisciplinary research: Methodologies for identification and assessment. Retrieve from https://www.mrc.ac.uk/documents/pdf/assessment-of-interdisciplinary-research/

Gowanlock, M., & Gazan, R. (2013). Assessing researcher interdisciplinarity: A case study of the University of Hawaii NASA Astrobiology Institute. *Scientometrics*, 94(1), 133-161.

Holland, G. A. (2008). Information science: an interdisciplinary effort? *Journal of Documentation*.

Huutoniemi, K., Klein, J. T., Bruun, H., & Hukkinen, J. (2010). Analyzing interdisciplinarity: Typology and indicators. *Research Policy*, *39*(1), 79–88.

Karlqvist, A. (1999). Going beyond disciplines: the meanings of interdisciplinarity. *Policy Sciences*, 32(4), 379-383.

Klein, J. T. (1990). *Interdisciplinarity: History, theory, and practice*. Wayne state university press.

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9), 1303–1319.

Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, *5*(1), 87–100.

Leydesdorff, L., de Moya-Anegón, F., & Guerrero-Bote, V. P. (2015). Journal maps, interactive overlays, and the measurement of interdisciplinarity on the basis of S copus data (1996–2012). *Journal of the Association for Information Science and Technology*, 66(5), 1001-1016.

Leydesdorff, L., Wagner, C., & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics*, *13*(1).

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS quarterly*, 35(2), 293-334.

MacLeod, M., & Nagatsu, M. (2016). Model coupling in resource economics: conditions for effective interdisciplinary collaboration. *Philosophy of Science*, *83*(3), 412–433.

Morillo, F., Bordons, M., & Gómez, I. (2001). An approach to interdisciplinarity through bibliometric indicators. *Scientometrics*, *51*(1), 203–222. https://doi.org/10.1023/A:1010529114941

Morillo, F., Bordons, M., & Gómez, I. (2003). Interdisciplinary in science: A tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology*, *54*(13), 1237–1249. https://doi.org/10.1002/asi.10326

Mugabushaka, A. M., Kyriakou, A., & Papazoglou, T. (2016). Bibliometric indicators of interdisciplinarity: the potential of the Leinster–Cobbold diversity indices to study disciplinary diversity. *Scientometrics*, *107*(2), 593-607.

Okamura, K. (2019). Interdisciplinarity revisited: evidence for research impact and dynamism. *Palgrave Communications*, *5*(1), 1–9. https://doi.org/10.1057/s41599-019-0352-4

Porter, A. L., & Chubin, D. E. (1985). An indicator of cross-disciplinary research. *Scientometrics*, *8*(3–4), 161–176. https://doi.org/10.1007/BF02016934

Porter, A. L., Cohen, A. S., D. Roessner, J., & Perreault, M. (2007). Measuring researcher interdisciplinarity. In *Scientometrics* (Vol. 72, Issue 1). https://doi.org/10.1007/s11192-007-1700-5

Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719-745.

Pratt, A. D. (1977). A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science*, *28*(5), 285-292.

Qin, J., Lancaster, F. W., & Allen, B. (1997). Types and levels of collaboration in interdisciplinary research in the sciences. *Journal of the American Society for information Science*, 48(10), 893-916.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, 82(2), 263–287.

Rhoten, D., & Pfirman, S. (2007). Women in interdisciplinary science: Exploring preferences and consequences. *Research policy*, 36(1), 56-75.

Rinia, E., van Leeuwen, T., & van Raan, A. (2002). Impact measures of interdisciplinary research in physics. *Scientometrics*, *53*(2), 241-248.

Schummer, J. (2004). Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics*, *59*(3), 425-465.

Soós, S., & Kampis, G. (2012). Beyond the basemap of science: mapping multiple structures in research portfolios: evidence from Hungary. *Scientometrics*, *93*(3), 869-891.

Steele, T. W., & Stier, J. C. (2000). The impact of interdisciplinary research in the environmental sciences: a forestry case study. *Journal of the American Society for Information Science*, *51*(5), 476. https://doi.org/10.1002/(sici)1097-4571(2000)51:5<476::aid-asi8>3.3.co;2-7

Wang, Q., & Schneider, J. W. (2020). Consistency and validity of interdisciplinarity measures. *Quantitative Science Studies*, *1*(1), 239–263. https://doi.org/10.1162/qss_a_00011

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, *5*(1), 14–26. https://doi.org/10.1016/j.joi.2010.06.004

Wang, J., Thijs, B., & Glänzel, W. (2015). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PloS One*, *10*(5).

Wang, X., Wang, Z., Huang, Y., Chen, Y., Zhang, Y., Ren, H., Li, R., & Pang, J. (2017). Measuring interdisciplinarity of a research system: detecting distinction between publication categories and citation categories. *Scientometrics*, *111*(3), 2023–2039. https://doi.org/10.1007/s11192-017-2348-4

Zhang, L., Rousseau, R., & Glänzel, W. (2015). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology*, 67(5), 1257-1265.

Zhang, W., Shi, S., Huang, X., Zhang, S., Yao, P., & Qiu, Y. (2020). The distinctiveness of author interdisciplinarity: A long-neglected issue in research on interdisciplinarity. *Journal of Information Science*. https://doi.org/10.1177/0165551520939499

Zhou, Q., Rousseau, R., Yang, L., Yue, T., & Yang, G. (2012). A general framework for describing diversity within systems and similarity between systems with applications in informetrics. *Scientometrics*, *93*(3), 787–812. https://doi.org/10.1007/s11192-012-0767-9

Zwanenburg, S. P., & Qureshi, I. (2019). Anticipating, avoiding, and alleviating measurement error: A synthesis of the literature with practical recommendations. *Australasian Journal of Information Systems*, 23.

Does Publicity in the Science Press Drive Citations?

Manolis Antonoyiannakis[1,2]

ma2529@columbia.edu

[1] Department of Applied Physics & Applied Mathematics, Columbia University, 500 W. 120th St., Mudd 200, New York, NY 10027 (USA)

[2]American Physical Society, Editorial Office, 1 Research Road, Ridge, NY 11961 (USA)

Introduction

Publishing is a selection process. Through a series of selections, journal editors decide which papers merit external review, which papers to send back to referees, and which papers to publish. But increasingly over the past 20 years, the selection of papers does not stop at publication, as editors curate lists of their favorite accepted or recently published papers, sometimes accompanying them with short summaries or longer commentaries by experts.

The community is paying attention. Researchers often promote (in their websites or resumés) their highlighted papers, while funding agencies track their grantees' progress by monitoring coverage in highlighting platforms.

Citation metrics agree well with peer review at the aggregate level (Traag & Waltman, 2019). So, does the post-acceptance "round" of review of highlighted papers produce a citation advantage? We address this question quantitatively, for several publicity markers on papers published in the journal Physical Review Letters (PRL) of the American Physical Society (APS). We thus extend our previous work (Antonoyiannakis, 2015).

Explanation of highlighting platforms

Cover image: Chosen for aesthetic reasons mainly.

Editors' Suggestions: Chosen for potential interest.

Viewpoints: Commentaries commissioned by the APS Physics editors and written by experts. They explain why a paper is important to the field.

Focus stories: Journalist-written news stories that explain the latest research to non-physicists.

Synopses: Short summaries of newsworthy results written by journalists and APS Physics staff.

Research Highlights: Short summaries of papers written by journal editors in the *Nature* journals.

News & Views: Commentaries by experts or (less often) by journal editors in the *Nature* journals.

Multiple Linear Regression

We perform Multiple Linear Regression on the citation data for papers in these highlighting platforms. As a first sample, we chose papers published from 2008–2012 (Figure 1). This amounts to 246 papers in the journal cover, 203 in Focus, 354 Viewpoints, 1232 Editors' Suggestions, 556 Synopses, 84 Research Highlights in Nature Physics, 36 News & Views in Nature Physics, and 91 Research Highlights in Nature. Clearly, the Viewpoint marker is the strongest predictor of citations. All else being equal, a Viewpoint marker allocates 10 additional citations to a PRL paper within 1 year of publication, and 44 citations within 5 years. The second strongest predictor of citations is Research Highlights in Nature. The third strongest predictors are Research Highlights in Nature Physics and Editors' Suggestions; and after that, Synopses and Covers**.** (The coefficients of News & Views in Nature Physics or Focus papers lack statistical significance.) We thus observe the following stratification of citation accrual:

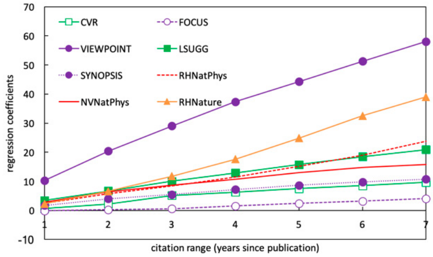


Figure 1. Coefficients of multiple linear regression for each highlighting platform.

$$\text{Viewpoint} > \text{RHNat} > \begin{cases} \text{Suggestion} \\ \text{RH NatPh} \\ \text{N\&V NatPh} \end{cases} > \begin{cases} \text{Synopsis} \\ \text{Cover} \\ \text{Focus} \end{cases} \quad (1)$$

Eq. (1) shows also a hierarchical pattern of decreasing scrutiny regarding importance during peer review. Viewpoints and Editors' Suggestions are internal highlights, for which the editors have the full benefit of peer review. Between these two, Viewpoints are vetted more, since they are discussed both among PRL journal editors and in a committee of Physics editors who seek further advice from experts on whether a Viewpoint is warranted. Synopses, Focus and Viewpoints are mutually exclusive, and since Viewpoints receive the highest level of scrutiny, Synopses and Focus articles are vetted less for importance. Research Highlights in Nature or Nature Physics, and News & Views in Nature Physics are all external highlights, which makes their selection more challenging, since their editors do not normally have access to the peer review files from PRL to aid their decisions. These journals are also more constrained in terms of space since their highlights cover several journals, and, for Nature, several fields. So, it makes sense that these platforms show a weaker prediction of citation accrual than Viewpoints, which receive the highest scrutiny among all platforms studied here. Finally, the difference in citation advantage between Research Highlights in Nature and Nature Physics may be because Nature is more selective than Nature Physics since it covers all fields of science.

The citation advantage of a Cover article, while statistically significant, is small. All else being equal, placement in the PRL cover adds no more than 1.5 additional citations per year. Thus, accidental or serendipitous publicity (recall that covers are chosen for aesthetics mainly) brings a small citation advantage, in direct analogy to the effect of visibility bias reported by Ginsparg and Haque (2009), for papers that accidentally end up in the top slot of the daily arXiv email listings. However, the citation advantage is clearly greater when publicity is deliberate and results from an endorsement of the paper's merit formed through peer review, as in a Viewpoint. So, publicity alone does help in terms of citations, but does not make a big difference unless it is supported by an endorsement, formed through peer review, that the paper has above-average merit.

Citation medians

We can also compare medians. In Figure 2 we show the annual median citation advantage for each highlighting platform. These results confirm our findings from multiple linear regression analysis.

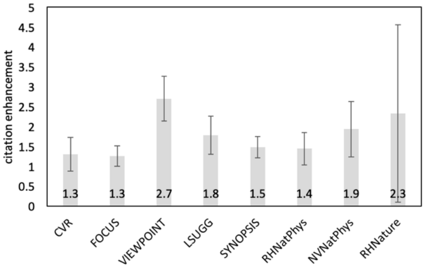


Figure 2. Annual enhancement of median citations per highlighting platform compared to the PRL journal. Publications from 2008–2018. Citation range 1–9 years after publication. Error bars at 99% confidence interval.

Clarivate list of highly cited papers

Can we use highlighting markers as predictors of a paper being *highly* cited? This question takes us beyond the previous analysis, since placement of a paper in a highly cited list is identifying extreme, not average, citation performance. As a benchmark for highly cited papers, we use the Highly Cited Papers (HCP) indicator of Clarivate Analytics. These papers are at the top 1% cited in their subject per year. We downloaded the list of HCP papers in physics published from 2010–2018, among which there are 1371 PRL papers. In Table 1 we show how well each platform predicts placement in the HCP list, i.e., how often highlighted papers are highly cited, which is the positive predictive value, or precision. The hierarchical pattern of Eq. (1) is thus reproduced. Again, we find that highlighting for importance correlates with citations—even for the extreme case of top-1% cited papers.

Table 1. Summary statistics for the precision (positive predictive value) of each platform in identifying highly cited papers.

Highlight	Count	Precision
CVR	446	0.099
FOCUS	378	0.063
VIEWPOINT	638	0.290
LSUGG	3269	0.143
SYNOPSIS	1117	0.117
RHNatPhys	116	0.164
NVNatPhys	33	0.121
RHNature	109	0.239

Conclusions

Our key conclusion is twofold. First, highlighting for importance identifies a citation advantage. Accidental or serendipitous publicity (i.e., mere visibility), gives a clearly smaller citation advantage. Second, the stratification of citations for highlighted papers follows the degree of vetting for importance during peer review. This implies that we can view the various highlighting platforms as predictors of citation accrual, with varying degrees of strength that reflect each platform's vetting level. More details are provided in Antonoyiannakis, 2021.

References

Antonoyiannakis, M. (2015). Editorial: Highlighting impact and the impact of highlighting: PRB Editors' Suggestions, *Physical Review B*, 92, 210001.

Antonoyiannakis, M. (2021). Does publicity in the science press drive citations? A vindication of peer review. In T. Vergoulis & Y. Manolopoulos (Eds.), *Predicting the Dynamics of Research Impact.* Springer (in press)

Ginsparg, P. & Haque, A. (2009). Positional Effects on Citation and Readership in arXiv, *Journal of the American Society for Information Science and Technology,* 60, 2203-2218.

Traag,V.A. & Waltman, L. (2019). Systematic analysis of agreement between metrics and peer review in the UK REF, *Palgrave Communications,* 5, 29.

Scientometrics of complexity: an institutional approach

Ricardo Arencibia-Jorge[1], José Luis Jiménez Andrade[2], Ibis A. Lozano-Díaz[3], Javier García-García[4] and Humberto Carrillo-Calvet[5]

[1] *ricardo.arencibia@c3.unam.mx*
Complexity Sciences Center (C3), National Autonomous University of Mexico (UNAM), Circuito Centro Cultural s/n, Coyoacan 04510, Mexico City (Mexico)

[2] *jlja@ciencias.unam.mx,* [3] *ibis.alozano@gmail.com*
Faculty of Sciences, National Autonomous University of Mexico (UNAM), Circuito Centro Cultural s/n, Coyoacan 04510, Mexico City (Mexico)

[4] *javier.garcia@c3.unam.mx,* [5] *humbertocarrillo@ciencias.unam.mx*
Complexity Sciences Center (C3) and Faculty of Sciences, National Autonomous University of Mexico (UNAM), Circuito Centro Cultural s/n, Coyoacan 04510, Mexico City (Mexico)

Introduction

Complexity sciences as a scientific field is an attractive domain that has not been deeply explored by scientometricians. Pollack, Alder & Sankaran (2014) offered evidence on Complexity Theory applied to research in Mathematics and Computer Science, and its growing application on management and organizational problems. Thomas & Zaytseva (2016) developed a more ambitious work, which analyzed approximately 1400 working papers published by Santa Fe Institute, describing the thematic evolution of Complexity research during the period 1989-2015, concluding that it is an intensely interdisciplinary field. The current work assumes the same institutional approach. The main aim is to characterize research developed by the Center for Complexity Sciences (C3) of the UNAM, according to the following questions: Which are the characteristics of mainstream research developed by C3? How does the international scientific community perceive this research? How multidisciplinary has this research been?

Materials and Method

Web of Science Core Collection (WoS) and InCites (supported by Clarivate Analytics), was used as the data source, covering ten years (2010-2019). Search strategies to identify entries for C3 and other eleven Complexity Sciences centers were developed, using the fields "Author Address" (AD) of the database. C3 output was characterized using bibliometric indicators provided by InCites, and multidisciplinary measures based on WoS subject categories (WCs) (Arencibia-Jorge, Vega-Almeida & Carrillo-Calvet, 2021). Self-Organizing Maps (SOM) were constructed using the software ViBlioSOM 2.0, developed at the UNAM. Key-words co-occurrence network was visualized in VOSviewer, developed at the Leiden University.

Results and Discussion

C3 research explores the Complex Systems, focused on strengthening links between science and society. Researchers at C3 usually came from other Faculties or Institutes at the UNAM, and they converge in C3 projects to develop interdisciplinary research. Their WoS output is in growth (432 documents; 13% of annual growth rate), with a remarked effort to publish their research results in high visible journals. H index was 29, showing averages of 8 citations per paper, 312 citations per year, and around 10 authors per article. Citing articles (2,466) came from 102 countries (headed by the United States 26.9%, China 11.9% and the United Kingdom 11.7%), published by 1,198 serials. Citing articles from Mexican institutions only constituted 14.5%. The evolution of C3's multidimensional performance profile (Figure 1) allowed us to confirm the strong publication activity during the last three years (dark/red zones in the graphs at the right side expressed intensively). The main efforts to publish articles in high visible journals (Q1) were observed during the first half of the decade, a period also characterized by an intense activity of international collaboration. The highest values of normalized impact were observed in 2019, which were related to a higher percentage of articles among the most cited publications (Top 10%).

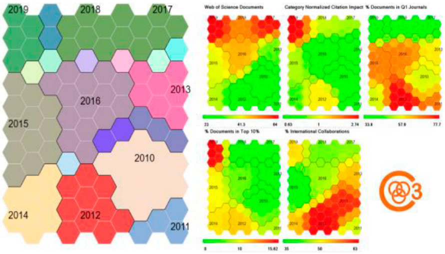


Figure 1. Evolution of C3's multidimensional performance profile (2010-2019).

The international collaboration of C3 (49.8%, with the participation of 98 countries from all continents) has been developed without avoiding the necessary links with Mexican institutions inside (22.4%) and outside the UNAM (27.8%). National collaboration was established with 56 institutions, particularly intense with the educative and health sectors.

The scientific production of C3 was published in 248 serials, covered by 74 research areas (48.7% of WoS research areas) and 105 WCs (41.2%). On the other hand, citing journals belong to 124 research areas (81.6%) and 196 WCs (76.9%). This confirms the diversity of research lines on which the C3 research is influencing. If we analyze the WCs core in articles (Thematic Concentration of production, TCp) and citing articles (Thematic Concentration of citations, TCc), researchers at the C3 published 80% of articles in 29 WCs, receiving 80% of citations from 37 WCs. This implies a Thematic Dispersion Index (TDI) of 32.8 (Figure 2). Compared with some of the most important international institutions dedicated to the study of complex systems, TDI of C3 was much higher than the rest.

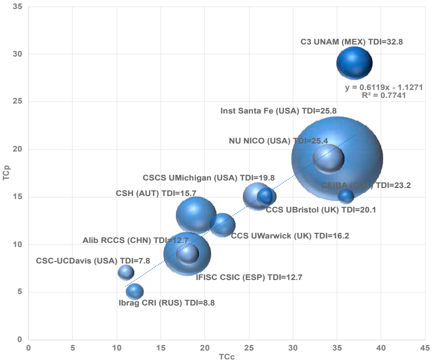
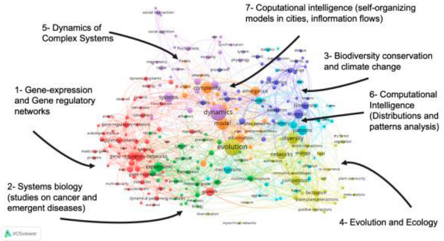


Figure 2. Multidisciplinarity of 12 international centers specialized in complex systems.

The keyword co-occurrence analysis confirmed that research's multidisciplinarity is a significant characteristic of C3, which is expressed by its seven main research fronts (Figure 3). The study of the dynamic of complex systems, gene regulatory networks, biology of systems, innovation, ecology, biodiversity or climate change, using computational intelligence and self-organizing models, with a set of socially-oriented challenges and goals, have been a characteristic of the Mexican institute, which requires a holistic vision and a quite ambitious strategy. This strategy was supported mainly by national funding (more than 50% funded by UNAM, 45.8% by national agencies, and only 30% by international agencies), a distinctive characteristic of C3 regarding the rest of studied international centers. C3 agreements and projects put the focus of C3 not only in mainstream science, but also in solving problems inherent to a country like Mexico, with serious health and mobility problems given its high population density, a huge social gap that generates the risk of internal crises (migration, corruption, violence), a rich ecosystem full of protected species and important natural resources, and a constant exposure to emergencies caused by hurricanes, earthquakes and volcanoes.



7- Coputational intelligence (self-organizing models in cities, inlformation flows)
5- Dynamics of Complex Systems
3- Biodiversity conservation and climate change
1- Gene-expression and Gene regulatory networks
6- Computational Intelligence (Distributions and patterns analysis)
2- Systems biology (studies on cancer and emergent diseases)
4- Evolution and Ecology
complexity
systems
dynamics
model
evolution
diversity
networks
gene regulatory networks
expression
mechanisms
competition
growth
VOSviewer

Figure 3. Research fronts identified in the C3 scientific output (2010-2019).

Conclusion

Since its foundation, C3 has promoted and integrated multidisciplinary academic groups to collaborate on large projects and face national challenges from the perspective of complex thinking. Scientists, artists, humanists and technicians were linked to C3 to address transdisciplinary challenges of national relevance, taking advantage of the synergy resulting from the interaction between different knowledge areas. During the period 2010-2019, C3 stood out for the growing output with expansive influences on the international scientific community, and an intense international and national collaboration activity. National funds for research were predominant. The multidisciplinary nature of research was evidenced by the diversity of topics addressed, as well as the variety of fields influenced by the published research.

Acknowledgments

This research was supported by the project "Scientometrics, Complexity and Science of Science", developed at UNAM-C3.

References

Arencibia-Jorge, R., Vega-Almeida, R. L & Carrillo-Calvet, H. (2021). A new thematic dispersion index to assess multidisciplinarity at different levels of aggregations. *In Proceedings of ISSI, in press.*

Pollack, J., Alder, D. & Sankaran, S. (2014) Mapping Complexity Theory: a Scientometric Approach. Emergence: *Complexity & Organization*, 16(2), 74-92.

Thomas, J. & Zaytseva, A. (2016). Mapping Complexity/Human Knowledge as a Complex Adaptive System. *Complexity*, 21(S2), 207-234.

Thematic multidimensional profile of Mexican scientific output in Dimensions 2010-2019

Ricardo Arencibia-Jorge[1], José Luis Jiménez-Andrade[2], Ibis A. Lozano-Díaz[3] and Humberto Carrillo-Calvet[4]

[1] *ricardo.arencibia@c3.unam.mx*
Complexity Sciences Center (C3), National Autonomous University of Mexico (UNAM), Circuito Centro Cultural s/n, Coyoacan 04510, Mexico City (Mexico)

[2] *ibis.alozano@gmail.com, {[3] jlja, [4] humbertocarrillo}@ciencias.unam.mx*
Faculty of Sciences and Complexity Sciences Center (C3), National Autonomous University of Mexico (UNAM), Circuito Centro Cultural s/n, Coyoacan 04510, Mexico City (Mexico)

Introduction

Mexican science has frequently been studied from a bibliometric approach (del Río, Russell, & Juárez, 2020; Lancho-Barrantes & Cantú-Ortiz, 2019; Villaseñor et al., 2017). Web of Science (Clarivate Analytics) and Scopus (Elsevier) are the main bibliographic databases used by the vast majority of these studies; and productivity, impact and collaboration have been the most studied variables. However, none of the previous papers have exploited the potential advantages of altmetric indicators; and there is no Mexican research that applies altmetric indices to research evaluation at the macro level using Dimensions, the last of the multidisciplinary bibliographic databases launched on the market. In this paper, Artificial Neural Network (ANN) and Self-Organizing Maps (SOM) are applied to create a scientometric profile, combining productivity, impact and altmetric measures to obtain a multidimensional representation of the Mexican scientific domains. The main aim is to characterize the Mexican scientific production in 22 subject categories covered by Dimensions.

Materials and Method

Dimensions, developed by *Digital Science & Research Solutions* (United Kingdom), was used as data source. Data were retrieved on October 8th, 2020. All the Mexican documents during the period 2010-2019 were identified, using the fields "Location-Research Organization" and "Publication Year". The field "Research categories" was analyzed through the Dimensions Analytics interface. In this field, the scientific output is structured in a classification scheme of 22 major fields and related sub-fields of research and emerging areas, based on the Australian and New Zealand Standard Research Classification (ANZSRC). A battery of indicators for each category were obtained: Publications (Npub), Citations (Ncit), Citations means (Ncit mean), Field Citation Ratio (FCR), Publications with attention (Pub Att) and Altmetric Attention Score (AAS), provided by Dimensions (https://www.dimensions.ai). Other three indicators were calculated: Activity Index (AI) (Frame, 1977); Attractivity Index (Attract) (Braun and Schubert, 1997); and the Relative Impact (RI), which was calculated to identify the relative impact of Mexico with respect to the world in each research category: Ncit mean $_{\text{Mexico (category)}}$ / Ncit mean $_{\text{World (category)}}$. The same principle was followed for other three Dimensions indicators (FCR, Pub Att and AAS). In all cases, values higher than 1 express a higher performance than the world. To facilitate the representation, a scale of values between -1 and 1 was used (Glänzel, 2000). The 0 value is the position of the world in each research category.

Tables in Microsoft Excel format with primary indicators relativized for each category were processed using a neural network technique. An artificial intelligence method was developed to automatically carry out the characterization of 22 main research categories in which Mexican scientific output is structured (Figure 1), obtaining a multidimensional scientometric map. The data mining procedure is based on the SOM family of neural networks (Teuvo Kohonen, 2013).

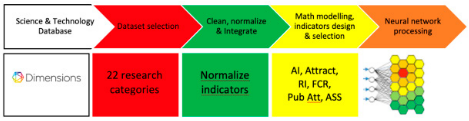

Science & Technology Database
Dataset selection
Clean, normalize & Integrate
Math modelling, indicators design & selection
Neural network processing
Dimensions
22 research categories
Normalize indicators
AI, Attract, RI, FCR, Pub Att, ASS

Figure 1. Methodology for data clustering and visualization using SOM neural networks.

The method was implemented in LabSOM, a program developed at the National Autonomous University of Mexico (UNAM). The SOM neural network was modeled as a two-dimensional hexagonal grid. Each hexagon represents an artificial neuron and, at the same time, a location where data points can be mapped. A nonlinear projection of data into the neural network was developed. During the neural network iterative training, the network learns to project similar patterns into close locations (hexagons) in the 2D map. Similarities between research area performances can be estimated by calculating the ''scientometric distance'' among their multidimensional representations (Villaseñor et al., 2017).

Results and Discussion

Standard bibliometric measures applied to the Mexican scientific output revealed a profuse activity in *Medical and Health Sciences* (24.7% of articles), which was also noted by Lancho-Barrantes & Cantú-Ortiz (2019). However, our study reveals that Mexican physicians' research has similar behavior to the world. The Mexican multidimensional thematic profile was used to compare the Mexican effort with the world ratio in each subject area (Figure 2).

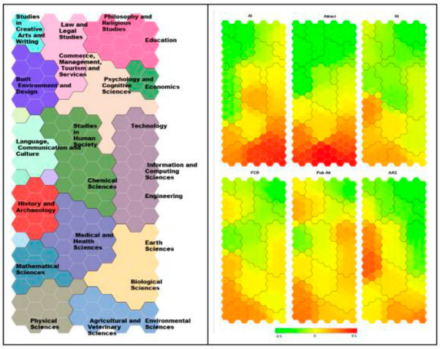


Figure 2. Thematic multidimensional profile of the Mexican scientific output during 2010-2019.

Subject areas were clustered according to the behavior of six indicators. The clusters with a darker color in the density graphs represent the areas in which Mexican scientific production stands out. The thematic areas located in the lower half of the graphs showed a better performance than the world in most of the indicators analyzed. *Agriculture and veterinary sciences* and *Environmental Sciences* experimented the best performance in almost all measures. *Physical sciences* and *Mathematical sciences* also showed values higher than the world, leading the FCR and Pub Att measures. *History and Archeology* achieved relevant citation-based indicators, such as the RI and the FCR, it was also the area with the highest AAS. *Biological Sciences* and *Earth Sciences* completed the number of areas with better national efforts. *Medical and Health Sciences* in Mexico showed values similar to the world; and some disciplines only highlighted using AI (*Chemical Sciences* and *Engineering*), RI (*Language, Communication and Culture*), FCR (*Built Environment and Design*), and Pub Att (*Technology* and *Information and Computing Sciences*).

The fact that Mexican research on *History and Archeology* has shown RI, FCR and AAS values considerably higher than the world average during the analyzed decade should not surprise anyone. Mexican history had strong links with world history before and after the Spanish conquest. Despite the lower AI (Mexican historians publish less than their world counterparts), the impact and altmetric measures expose high-quality Mexican research, relevant to the media. Mexican researchers on *Environmental Sciences* and *Agricultural and Veterinary Sciences* also showed a remarked performance. Particularly, in *Environmental Sciences,* our results demonstrated the Mexican efforts to solve the national environmental problems during the decade. Physics and Mathematics are specialties where Mexico has developed relevant research, which confirmed previous bibliometric reports (del Río, Russell & Juárez, 2020). However, with the solitary exception of *History and Archeology*, humanities and social sciences still showed low performances.

Conclusion

SOM-based science mapping allowed a better comprehension of Mexican research areas during the last decade. *Agriculture and veterinary sciences*, *Environmental Sciences*, *Physical sciences*, *Biological Sciences*, *Earth Sciences*, *Mathematical Sciences* and *History and Archeology* achieved the best bibliometric performances in relation to the world. Altmetric indicators offered an interesting approach that would be seriously analyzed for future research evaluation policies.

Acknowledgments

The authors thank Dr. Javier García-García, and the free data access program for scientometricians developed by *Digital Science & Research Solutions.*

References

Braun, T. & Schubert, A. (1997). Dimensions of scientometric indicator datafiles - World science in 1990-1994. *Scientometrics*, 38, 175-204.

del Río, J. A., Russell, J. M. & Juárez, D. (2020). Applied physics in Mexico: mining the past to predict the future. *Scientometrics*, 125, 187-212.

Frame, J. D. (1977). Mainstream research in Latin America and the Caribbean. *Interciencia*, 2, 143-148.

Glänzel, W. (2000). Science in Scandinavia: a bibliometric approach. *Scientometrics*, 48, 121-150.

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural networks*, 37, 52-65.

Lancho-Barrantes, B. S. & Cantú-Ortiz, F. J. (2019). Science in Mexico: a bibliometric analysis. *Scientometrics*, 118(2), 499-517.

Villaseñor, E. A., Carrillo-Calvet, H. & Arencibia-Jorge, R. (2017). Multiparametric characterization of scientometric performance profiles assisted by neural networks: a study of Mexican higher education institutions. *Scientometrics*, 110(1), 77-104.

A new thematic dispersion index to assess multidisciplinarity at different levels of aggregations

Ricardo Arencibia-Jorge[1], Rosa Lidia Vega-Almeida[2] and Humberto Carrillo-Calvet[3]

[1] *ricardo.arencibia@c3.unam.mx*
Complexity Sciences Center (C3), National Autonomous University of Mexico (UNAM), Circuito Centro Cultural s/n, Coyoacan 04510, Mexico City (Mexico)

[2] *vegaalmeida.rosa@gmail.com*
Information Science PhD Program, University of Havana, Havana (Cuba)

[3] *humbertocarrillo@ciencias.unam.mx*
Faculty of Sciences and Complexity Sciences Center (C3), National Autonomous University of Mexico (UNAM), Circuito Centro Cultural s/n, Coyoacan 04510, Mexico City (Mexico)

Introduction

Multidisciplinary research deals with the work of specialists from different fields in a disciplinary context, without cross its boundaries. If, in this process, researchers integrate knowledge and methods from their disciplines in a new synthetic, coordinated and coherent approach, an interdisciplinary domain emerges. There is a fuzzy line among both contexts, which still is a challenge for scientometricians. Usually, and not without criticism, researchers use a top-down approach dependent on existing classifications schemes to create a knowledge domain's taxonomy (Moschini et al., 2020). This approach assumes that assigning journals to multiple subject categories can be considered to analyze interdisciplinarity (Rousseau, Zhang & Hu, 2019). The study of communication between authors, institutions, documents or citations in these complex networks illustrates the diversity of research areas (Porter & Rafols, 2009; Wagner et al., 2011). This paper uses the Web of Science (WoS) list of subject categories (WCs) as a proxy to measure the multidisciplinary scope of knowledge domains, research institutions and individuals. The main aim is to show a new set of WCs-based bibliometric indicators, which takes into account documents/disciplines and citing documents/influenced disciplines in a new approach to measure multidisciplinarity, demonstrating how elements from different disciplines (and probably inter- or cross-disciplinarity concepts inherent to it) are present in the scientific output of entities and in their spheres of influence.

Materials and Method

WoS Core Collection was used as the data source. We developed two indicators to analyze thematic concentration (minimum number of WCs involved) of publications and citations received by each entity (researcher, institution or knowledge domain). Once the scientific production was retrieved, datasets were refined using WCs, following a descending order, until retrieving a number of documents closer to 80% of the total output. The number of refined WCs is the thematic core that involves the entities' scientific production (Thematic concentration of production, TCp). After that, using the option "citation inform", the total amount of citing documents from each entity was retrieved. Following the previously exposed principle, the set of cited documents was refined using WCs, until retrieving a number closer to 80%. In this case, the number of refined WCs is the thematic core that involves documents influenced by entities (Thematic concentration of citations, TCc).

Setting 80% as threshold reduced size-dependence limitations. The assumption that not necessarily a high productive entity, with a wide range of topics studied, implies a high number of WCs in the thematic core, deals with a dichotomy specialization/multidisciplinarity which is quite presented in all cases studied at the macro, meso and micro level. Based on the thematic concentration of documents and citing documents, a new Thematic Dispersion Index (TDI) was calculated, to assess the multidisciplinary scope of research. This measure was calculated using the following formulation:

$$TDI = \sqrt{TCp \times TCc}$$

Values closer to 1 express a high level of specialization or disciplinary concentration. Higher values will determine, in a growing way, the multidisciplinary scope of research. At macro level, we organized our units of analysis in three datasets, according to the following search strategies: a) TS="Solar flare*"; b) TS="Ethnomethodol*"; c) TS="Artificial intelligence", during the period 1970-2019. Results at meso level (eight research institutes belonging to the UNAM) and micro level (Members of the researcher staff of the Complexity Sciences Center (C3) at the UNAM) were also discussed.

Results and Discussion

Although each macro-level dataset showed exponential growth of scientific output ($R^2 > 0.9$) and linear growth of WCs involved ($R^2 > 0.9$), TCp, TCc and TDI calculated for the 5-year stages during the whole period showed a different behavior (Figure 1).

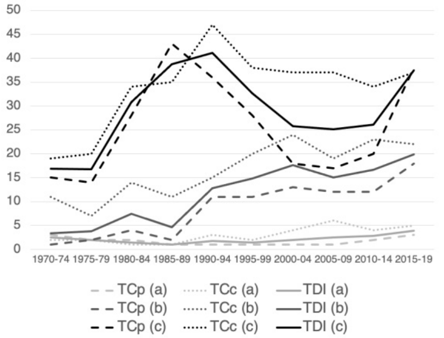


Figure 1. Multidisciplinarity measures applied to a) solar flares, b) ethnomethodology, and c) Artificial Intelligence.

Research on "solar flares" was highly productive (293 to 1839 documents per stage) but showed the classic disciplinary behavior (Astronomy and Astrophysics), with less than five WCs involved in 80% of papers and citing papers. Research on ethnomethodology was less productive (14 to 626). Still, a growing trend was observed in the WCs core during the whole period, putting in evidence the gradual transition from a very specialized topic of Sociology (1970-1989) to a transversal methodology used by linguists, psychologists, physicians and professionals in the fields of management, communication, business, computer science, or even Library and Information Science (LIS). Finally, documents on "Artificial Intelligence" (60 to 24828 documents per stage) showed a seesaw behavior, which could be related to epistemological highlights within the field. After an initial expansion of topics studied (1970-1994), research evolved to a "paradigmatic" phase of development (1995-2009), and then to a re-configuration of the discipline as a technoscience (2010-2019), involving an explosion of technological solutions for productive systems (Arencibia-Jorge et al. Manuscript under review).

At the institutional level, multidisciplinary patterns in research activities from eight research institutes of the UNAM were identified. Two of them stood out above the others: C3 (TDI = 32.8) and the *Instituto de Investigaciones sobre Matemáticas Aplicadas y Sistemas* (IIMAS, TDI = 32.9). In the first case, the study of the dynamic of complex systems, gene regulatory networks, biology of systems, innovation, ecology, biodiversity or climate change, using computational intelligence and self-organizing models, requires a holistic vision that can only be possible through the adoption of a multidisciplinary perspective. In the second case, the transversality of research is an implicit characteristic of the institute. At the micro level, C3 high productive senior researchers were studied, which include Physicists (40%), Biologists (22%), Physicians (12,5%), mathematicians, engineers, chemists, one computer scientist and one information scientist. The diversity of the staff composition and the complex nature of the studied problems were clearly revealed by TDI values, which were higher than 5 in 69% of cases and even higher than 10 in the 34% of cases. Therefore, even at the individual level TDI offered optimal values to assess multidisciplinary research.

Conclusion

The proposed WCs-based measures helped determine the multidisciplinary scope of research at different aggregation levels. Despite the limitations of classification schemes, the new approach used the Pareto principle to obtain a less size-dependent WCs core (not affected by productivity or impact), which allowed characterizing the thematic dispersion of research developed by entities at the macro, meso and micro level. In each of the analyzed datasets, 80% of published documents and citing articles represented the main research lines produced and influenced by an entity. This basic approach could be complemented with diversity measures, which could identify the levels of knowledge integration in the research fronts in greater depth.

Acknowledgments

This research was supported by the UNAM-C3. We appreciate the collaboration of Javier García-García.

References

Moschini, U., Fenialdi, E., Daraio, C. Ruocco, G. & Molinari, E. (2020). A comparison of three multidisciplinarity indices based on the diversity of Scopus subject areas of authors' documents, their bibliography and their citing papers. *Scientometrics,* 125(2), 1145-1158.

Porter, A. & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics,* 81(3), 719-745.

Rousseau, R., Zhang, L. & Hu, X. (2019). Knowledge integration: Its meaning and measurement. In *Springer handbook of science and technology indicators* (pp. 69-94). Springer, Cham.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., et al. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics,* 5(1), 14-26.

‘Larks’ and ‘Owls’: More on Eliciting Human Moods and Chronotypes from Psychometric Variables in Twitter Data

Aparna Basu

aparnabasu.dr@gmail.com
Independent Researcher, Formerly Emerita at CSIR-NISTADS and
Guest Faculty, South Asian University, New Delhi (India)

Introduction

Of late, there is increasing interest in Altmetrics, that is, measures of informal mentions of research papers in social media, such as blogs, Facebook, news, Twitter etc. These entities are like citations to scientific papers, but have distinct characteristics. In twitter data one can look at citation contexts or content words. This can give a clue to the mood of the writer (positive or negative). Unlike citations, tweets are subject to variations in time. We re-analyse here the diurnal variation of a sample of tweets studied by Dzogang et al. (2018) and its interpretation in terms of human moods. According to the authors, the variation in the mood of the tweets showed that people were more positive in the mornings, but had more depressive thoughts at night. Our objective is to show that one can re-interpret the same data and analysis to obtain a completely different result. According to our interpretation, human beings can be classed as early risers or ‘larks’ and late risers or ‘owls’. Positive mood is more strongly associated to the ‘larks’ and Negative emotions to the ‘owls’. This makes interesting inroads into psychology and sleep patterns in humans, bringing into the argument the evolutionary reasons for varying sleep patterns.

Background, Data and Methods

Data

Dzogang, Lightman and Cristianini (2018) performed a large scale experiment with 800 million tweets from urban centres in the UK, collecting data over 4 years. Tweets were anonymised. The data was sampled every hour during the day-night cycle. A fairly stable diurnal (day and night) variation of tweet intensity was observed. It was also observed that the all words were not evenly distributed in time.

The analysis was performed in 3 stages. (1) Data collection and anonymisation, (2) Text Analysis/ Sentiment Analysis and (3) Factor Analysis.

Text Analysis/ Sentiment Analysis

In this stage, the words in the tweets were linked to the corresponding ‘mood’ that they expressed, based on a standard source, the text analysis program Linguistic Inquiry and Word Count. Words of positive affect were linked to emotions such as enthusiasm, delight, activeness and alertness – or of negative affect - like fear, guilt, anger, disgust. This process is also known as sentiment analysis

Factor Analysis

Factor analysis was performed on the ‘mood’ associated with the tweet words and their diurnal variation was noted. It brought out two principal factors F1 and F2 in the data which showed distinct diurnal variation, and explained 85% of the variability in the original data. Moreover, they loaded differently on the mood variables. (Figure 1). Dzogang et al. do not offer any interpretation for the factors F1 and F2.

Dzogang’s conclusions

Dzogang et al. state, “Overall, we see strong evidence that our language changes dramatically between night and day, reflecting changes in our concerns and underlying cognitive and emotional processes.”

It implies that the same person instinctively expresses different emotions during different times. What is more, all persons follow the same pattern. Human emotions vary in a cyclical way according to the circadian rhythm, being more positive in the morning and more negative towards evening.

Sleep, ‘Larks’ ‘Owls’: The Sentinel Hypothesis

In the entire discussion, no account seems to have been taken of sleep time when at least part of the population would be inactive. What is missing is the proper place and role of sleep.

Sleep is essential for survival, yet, for animals, it also represents a time of extreme vulnerability to predation and environmental dangers. The sentinel hypothesis proposes that group-living animals share the task of vigilance during sleep, with some individuals sleeping while others are awake. Asynchronous sleeping habits among members of a group, confers an evolutionary advantage, In modern populations asynchronous sleep or the existence of chronotypes may represent a legacy of natural selection acting to reduce the dangers of sleep (Samson, 2017; Beauchamp, 2015; Randler, 2014).

The Sentinel Hypothesis was originally proposed by Frederick Snyder. According to him, “*Man and other animals have learned that under conditions of danger it is safe to sleep only if sentinels ate*” (Snyder, 1966) The tendency to have asynchronous sleep patterns among individuals has been noticed in present human society as well as a legacy of our evolutionary past. Those who rise early, or go to bed late, have been typified by the term chronotype, or Larks and Owls respectively.

Our Interpretation

The factors F1 and F2 discussed earlier represent only the active part of the population and not the part that is asleep. F1 is shifted from F2 by a few hours on the time axis (Figure 1). From the sleep times, we can infer that F1 will correspond to early risers, the Larks, and F2 to the late risers or Owls (see Figure 1). The following description shows that one could also associate cognitive and emotional traits to the Larks and Owls based on the experiment.

"*The first had peak expression time starting at 5am/6am, it correlated with measures of analytical thinking, with the language of drive (e.g power, and achievement), and personal concerns. It is anti-correlated with the language of negative affect and social concerns. The second factor has peak expression time starting at 3am/ 4am, it correlates with the language of existential concerns, and anti-correlates with expression of positive emotions*". (Dzogang et al., 2018)

Discussion

We give an alternative interpretation to the diurnal variation of tweet words in the experiment (Dzogang et al., 2018) by including the role of sleep as a constraint in sending out tweets. We see that tweet data can be used in different ways than just mapping trends or as a proxy for citation. The present example includes sentiment analysis to deal with the psychology of tweeterss. Some advantages of using tweet data is the massive data size, but even more importantly that it is non-invasive, in real time, and does not require people to recollect details as in the case of data collection with questionnaires.

Our re-interpretation of Dzogang's experiment with 800 million tweets gives empirical evidence that we have a binary population of humans with different cognitive characteristics and emotional behaviours, and distinguished by their sleep characteristics, viz. the larks and the Owls.

One might end with just one word of caution. Much of the vocabulary associated with behaviour, cognition or emotion are value loaded (say, positive and negative affect) and one could easily have a situation where people are discriminated against for some intrinsic qualities left over from evolution. The present working day and working hours and its demands, for example, would be quite unsuitable for Owls. The same would be true for Larks on night shifts.

Acknowledgement

I thank the authors F. Dzogang, S. Lightman, and N. Cristianini, for use of their figures.

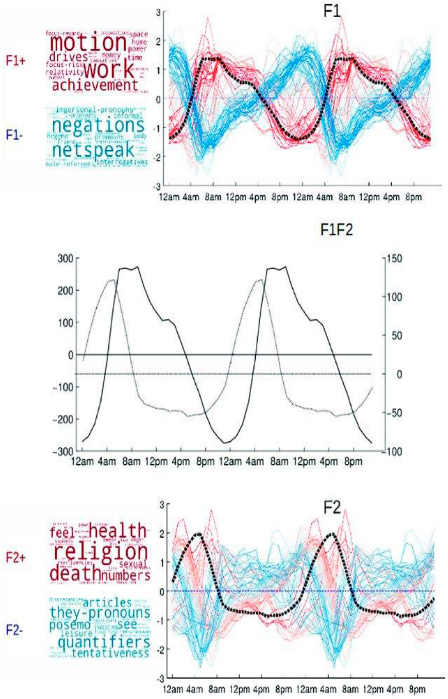


Figure 1. Diurnal variation of F1 and F2 , and words of +ve and –ve affect.

*Recreation from (*Dzogang et al., 2018)

References

Beauchamp, G. (2015). *Animal vigilance: monitoring predators and competitors*. Quebec, Canada: Academic Press.

Dzogang, F., Lightman, S. & Cristianini, N. (2018). Diurnal variations of psychometric indicators in Twitter content. *PLoS ONE*, 13(6), e0197002.

Randler, C. (2014). Sleep, sleep timing and chronotype in animal behaviour. *Anim. Behav*., 94, 161-166.

Samson, D. R., Crittenden, A. N., Mabulla, I. A., Mabulla, A. Z. P. & Nunn, C. L. (2017). Chronotype variation drives night-time sentinel-like behaviour in hunter – gatherers. *Proc. R. Soc.* B, 284, 20170967.

Golder, S. A. & Macy, M. W. (2011). Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*, 333(6051), 1878-1881

Snyder, F. (1966). Toward an evolutionary theory of dreaming. *Am. J. Psychiatry*, 123, 121-136. (doi:10. 1176/ajp.123.2.121)

Wulff, K., Gatti, S., Wettstein, J. G. & Foster, R. G. (2010). Sleep and Circadian Rhythm disruption in psychiatric and neurodegenerative disease, *Nature Reviews*, www.nature.com.

Robustness of citation networks retrieved from queries

Alexandre Benatti[1], Henrique Ferraz de Arruda[1], César Henrique Comin[2], Filipi Nascimento Silva[3] and Luciano da Fontoura Costa[1]

[1] *alexandre.benatti@usp.br, h.f.arruda@gmail.com, and ldfcosta@gmail.com*
FCM, Institute of Physics of São Carlos, Avenida Trabalhador São-carlense, nº 400, Parque Arnold Schimidt - CEP 13566-590, São Carlos - São Paulo (Brasil)

[2] *chcomin@gmail.com*
Department of Computer Science, Federal University of São Carlos Rodovia Washington Luís km 235, CEP 13565-905, São Carlos - São Paulo (Brasil)

[3] *filipinascimento@gmail.com*
Indiana University Network Science Institute, Bloomington, IN, 47408 (USA)

Introduction

Citation networks is a central element in Science of Science (Fortunato et al., 2018). Methods to construct and analyze such networks play a pivotal role in understanding how science is organized and unfolds. With the advent of freely available and comprehensive scholarly datasets, comes the need for techniques to select and extract subsets of records that are particularly meaningful. A common example of that is building citation networks from records associated with a specific knowledge area. One way to accomplish that is by matching publications titles and abstracts against a set of query terms.

Among the most interesting characteristics of citation networks are their community structure. In Silva et al., 2016, community detection was employed to build a hierarchy of topics that can be used to summarize a subset of the scientific literature. This technique was further improved in Ceribeli, de Arruda & Costa (2019), in which a multiscale representation of the communities was considered. However, the choice of the query terms may directly impact the constructed network and consequently its community structure. For instance, suppose a community *A* in a citation network acts as a bridge between two other communities *B* and *C*. In the case such a community has only records recovered from a single search term, B and *C* can become completely disconnected. In other cases, the terms may have a reasonable number of overlapping records that may lead to no significant changes when terms are omitted.

In this paper, we propose an approach to better understand the effects of the initial selection of query terms. In particular, we explore how the community structure of a citation network is affected when some query terms are omitted from the data retrieval and network construction steps. Furthermore, here we test the methodology for a large citation network obtained from the *Microsoft Academic Graph* (MAG) (Sinha et al., 2015) by using a set of query terms related to Machine learning. Preliminary results indicate that the community structure for this network is robust to changes in the set of query terms.

Employed network

As described in (Silva et al., 2016), a network for a scientific field is retrieved from records that match (or contains) a pre-defined set of search terms (keywords), which is employed to limit a given scope to the desired field. Here, we retrieve all documents from the MAG (obtained in 2020) that contain at least one of the keywords in its title or abstract. To create the network, each retrieved document is considered as a node, and the directed edges are created from the citations between them. So as to eliminate undesired documents, only the weakly major connected component is considered. One example of the employed network is illustrated in Figure 1.

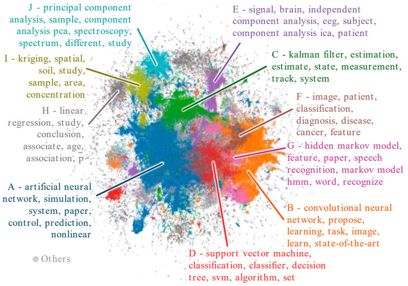


Figure 1. Visualization of the network nodes, where the positions were defined by node2vec and UMAP projection. The detected communities are represented by colors.

Employed methodology

By considering the network described in the previous section, the communities are detected. As in (Ceribeli, de Arruda & Costa, 2019), we use the *Infomap* algorithm (Rosvall, Axelsson & Bergstrom, 2009), which is often considered in science of science applications. From the detected communities, we obtained the subjects (Silva et al., 2016).

In order to test the robustness of the employed methodology, we compare the network obtained from the entire and reduced sets of keywords. The latter type is henceforth called *modified network*. Because the subject identification approach is based on the community structures, we compare the communities identified in the complete and modified networks. More specifically, for each comparison, the Normalized Mutual Information (NMI) is calculated.

Results and discussion

We illustrate our methodology by considering a set of keywords in the field of artificial intelligence. Fig. 1 shows the obtained network, in which colors represent communities. The identified subjects are also shown. Here, the modified versions consist of networks obtained with reduced sets of keywords. For the sake of simplicity, each of the reduced sets corresponds to the entire set minus a single keyword. Documents containing a removed keyword kept in the network if they contain another keyword in the reduced set. Fig. 2 illustrates the obtained NMI and the number of vertices of the networks normalized by the complete network size.

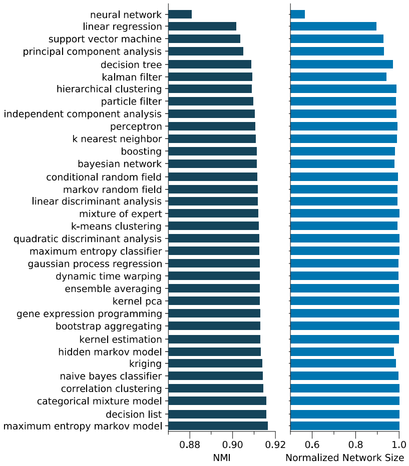


Figure 2. NMI and the normalized network sizes of the comparison between community organizations of modified and complete networks.

The results were found to indicate good stability in terms of the networks' community organization. By comparing the standard deviation of the NMIs (0.006) and the normalized sizes of the modified networks (0.075), we found that the variation of NMI is the lowest. The worst-case scenario was NMI = 0.88 (removal of the keyword "neural network"). This case also corresponds to the smallest obtained network, with 508.027 nodes (56,6% of the complete network size). Even with a drastic reduction of the network size, the community structure was remained stable, as indicated by the high value of NMI.

Conclusions

A citation network was obtained by considering the MAG dataset and a given set of machine learning keywords. The organization of communities was obtained, and the respective subjects were identified. To study the robustness of this methodology, we removed keywords from the original set and compared the original with the obtained networks. The different sizes of the modified networks reveal that the deletion of some keywords can remove a substantial number of scientific documents. However, for the remaining documents, the original community structure tends to be preserved. These results illustrate that missing query terms used to delineate a field may not have huge impact to some network analyses.

Acknowledgments

AB thanks Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. HFA thanks FAPESP (grant #2018/10489-0 and #2019/16223-5) for support. LFC thanks CNPq (grant no. 307085/2018-0) and NAP-PRP-USP for sponsorship. CHC thanks FAPESP (grant #2018/09125-4) for support. This work has been supported also by FAPESP grant #2015/22308-2. This material is partially supported by AFOSR #FA9550-19-1-0391. Research carried out using the computational resources of the Center for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP (grant 2013/07375-0).

References

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... & Barabási, A. L. (2018). Science of science. *Science*, 359(6379).

Silva, F. N., Amancio, D. R., Bardosova, M., Costa, L. da F. & Oliveira Jr, O. N. (2016). Using network science and text analytics to produce surveys in a scientific topic. *Journal of Informetrics*, 10(2), 487-502.

Ceribeli, C., de Arruda, H. F. & da F. Costa, L. (2021). How coupled are capillary electrophoresis and mass spectrometry? *Scientometrics*, 1-11.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B. J. & Wang, K. (2015). *An overview of microsoft academic service (mas) and applications*. In Proceedings of the 24th international conference on world wide web (pp. 243-246).

Rosvall, M., Axelsson, D. & Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, 178(1), 13-23.

Extracting Sentences Concerning Research Results of Citations from Full Text of JASIST

Jiaxin Bian[1], Si Shen[2] and Dongbo Wang[3]

{[1] 120107222870, [2] shensi}@njust.edu.cn
Nanjing University of Science and Technology (China)

[3] db.wang@njau.edu.cn
Nanjing Agricultural University (China)

Introduction

The related work part of an academic paper usually contains a summary of previous studies. These sentences generally appear in the second part of academic papers, and are of great significance for quickly understanding research results in a specific field and discovering research hotspots. However, considering the sparseness of the sentences related to the research results of citations and the time-consuming of reading the literature, it is particularly important to extract and analyze the sentences related to the research results of citations. So far, the research on the extraction of academic full text content has only reached the stage of using regular expression for the extraction. For example, in the work of Wang and Zhang (2020) an algorithm dictionary based on manually annotated paper content was constructed, and sentences containing algorithms were extracted through dictionary-based matching methods, and the impact of algorithms mentioned in academic texts was analyzed. In order to fill this gap, this research adopts a deep learning method to extract sentences related to the research results of citations from the academic full text, and verifies the effectiveness of this method. Then in order to make deeper research at these sentences, clustering experiments were carried out on the academic full text and three types of sentences concerning the research results were identified.

Data Set and Method

Data were collected from full-text articles from JASIST (2017-2020). From 2017 to 2020, a total of 502 papers were retrieved. We annotated the experimental results and conclusions in the citation sentences of academic texts. For example, the sentence "In Egghe and Rousseau (In press), we proved the following results for the moveout Lotka function, extending similar results for the Classical case." was labeled as the research results of citations. The labelling tool is the BRAT platform. After annotation, there are 6,083 sentences concerning research results of citations in these papers. And the specific information of the corpus is shown in Table 1.

Table 1. Basic Information of Corpus.

Num.	Type	Count
1	Papers	502
2	Total Sentences	138,197
3	Research Results sentences	6,083
4	Marked sentences in each article average number	12.12
5	Average words number in each sentences	31.88

The BERT model used in this paper is a pre-training model, which is implemented based on a bi-directional Transformer Encoder (Devlin et al., 2018). Our experiment was trained on the BERT-base Cased version.

In this paper, NVIDIA Tesla P40 graphics processor (GPU) was used for neural network training. The performance parameters of the computer used in the experiment are as follows:

Table 2. Basic Information of Experimental Environment.

Hardware Name	Performance Parameter
CPU	48 Intel(R) Xeon(R) CPU E5-2650 v4@2.20GHz
Memory	256GB
GPU	6 PCS NVIDIA Tesla P40
Video Memory	24GB
Operating System	CentOS 3.10.0

Results

In the training of BERT model, the pre-training model with 12-layer, 768-hidden and 12-heads was chosen, and then 10-fold cross-validation tests were performed, and evaluated by P, R and F values. The evaluation results are shown in Table 3.

It can be seen from Table 3 that the maximum Macro-AVG F-value of the Bert model classification result is 75.11%, which is lower than 61.02%, the Macro-AVG F-value of the SVM model we used before. The experimental results of the SVM model are shown in Table 4.

Table 3. Results of 10-Fold Cross-Validation on BERT.

Num.	Research Results Sentences F-value	Non-Research Results Sentence F-value	micro-AVG F-value	Macro-AVG F-value
1	43.63%	92.93%	87.44%	70.17%
2	54.88%	95.03%	91.05%	75.11%
3	38.54%	95.25%	91.18%	66.91%
4	50.80%	94.55%	90.19%	73.41%
5	49.80	95.11%	91.08%	72.45%
6	40.6%	93.92%	88.97%	67.26%
7	37.98%	93.37%	88.02%	67.19%
8	43.60%	94.32%	89.68%	70.13%
9	47.24%	94.55%	90.12%	71.39%
10	41.62%	93.06%	87.59%	71.22%
MAX-F	54.88%	95.25%	91.18%	75.11%
AVG-F	44.87%	94.21%	89.53%	70.52%

Table 4. Results of 10-Fold Cross-Validation on SVM.

Num.	Research Results Sentences F-value	Non-Research Results Sentence F-value	Macro-AVG F-value
1	26.17%	95.88%	61.02%
2	13.94%	94.45%	54.19%
3	9.41%	96.05%	52.72%
4	22.95%	95.83%	59.38%
5	17.92%	95.58%	56.74%
6	13.52%	94.87%	54.19%
7	17.73%	96.18%	56.95%
8	17.13%	96.17%	56.65%
9	16.54%	95.76%	56.14%
10	9.66%	89.20%	50.83%
MAX-F	26.17%	96.18%	61.02%
AVG-F	15.50%	95.00%	55.88%

A comparison of Table 3 and Table 4 shows that the training effect of the BERT model is significantly better than that of the SVM model, which proves the applicability of the BERT model in extracting sentences concerning the research results of citations. Then, in order to further mine the sentence information related to the research results of citations, the K-Means model was used to cluster the sentences. The result is shown in Figure 1.

As can be seen from the figure above, the research result sentences of citations extracted from the academic full text are divided into 3 types, each of which is represented by different colours. The emotion of citing is often the judgment of subjective emotion, which can reveal the author's attitude towards the citation. Therefore, on the basis of the above clustering experiment, this paper divides the sentences into support, opposition and neutral according to their emotional tendencies.

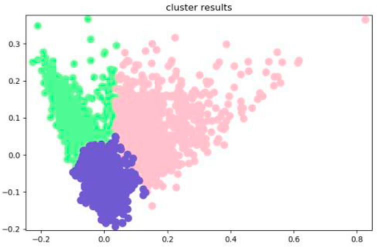


Figure 1. Graph of sentence clusters in JASIST academic full texts.

Conclusion

This paper explores the academic full text corpus. The BERT model was used to automatically extract the sentences concerning research results of citations, and the maximum harmonic mean reached 75.11%. Compared with the extraction effect of the previously used SVM model, it proves the superiority of the BERT model in the automatic sentence extraction task of the citation research results. Then, on the semantic level, K-means clustering was used to divide these sentences into three types according to their emotional tendencies, namely, support, opposition and neutral. This research filled the gaps in the research of methods to extract sentences concerning research results of citations from the academic full text, which provides a basis for further classification of such sentences.

Acknowledgments

The authors acknowledge the National Natural Science Foundation of China (Grant Numbers: 71974094) for financial support.

References

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: 1810. 04805v1*

Wang, Y. & Zhang, C. (2020). Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. *Journal of Informetrics*, 14, 4.

Shifting the ARWU methodology to Q1 journals: what are the consequences for lower-tier universities?

Srđa Bjeladinović[1], Veljko Jeremić[2] and Mladen Stamenković[3]

{[1]srdja.bjeladinovic, [2] veljko.jeremic}@fon.bg.ac.rs
University of Belgrade - Faculty of Organizational Sciences, Jove Ilića 154, 11000, Belgrade (Serbia)

[3] mladen@ekof.bg.ac.rs
University of Belgrade - Faculty of Economics, Kamenička 6, 11000, Belgrade (Serbia)

Introduction

Academic Ranking of World Universities (ARWU) represents one of the most renowned universities' ranking lists (Safón, 2019). Throughout the years, it has been revised to a minor extent to address some of the critiques (Mussard & James, 2018). The majority of effort from the ARWU creators were oriented towards diversification of the ranking lists offered. Thus, several lists emerged, such as the Global Ranking of Sport Science Schools and Departments, ARWU-FIELD and Global Ranking of Academic Subjects (ARWU, 2021). The ultimate one gained much of the attention, with 54 subjects presented in the 2020 edition. The major shift in methodology consisted of including solely Q1 (ranked in the first quarter according to their respective JCR category - Sorz et al., 2020) journals in the 2020 edition. The shift in methodology catalysed the rapid decline of ranks for numerous lower-tier universities, including the University of Belgrade (UB). In the 2018 edition, the UB was among the best universities in as many as 27 of the 54 scientific research areas. In 2020, the change in methodology had led to the University of Belgrade being positioned within 12 scientific research areas (Jeremić & Stamenković, 2020). Having this in mind, we wanted to explore the effects on lower-tier universities when potentially switching PUB score (based on Articles published in SCIe and SSCI journals) to Q1 score (taking into account solely Q1 journals) in major ARWU Rankings published each August 15th.

Methods and Data

As a case study, we obtained the data containing WoS indexed Articles (SCIe and SSCI indexed journals, Q1 journals, year of publication 2019) published by researchers from the selected universities. In total, ten universities are selected, seven universities (*University of Belgrade*, three universities from China: *China University of Mining and Technology – Xuzhou*, *Lanzhou University* and *Nanjing Medical University*, and three universities from Brazil: *Federal University of Minas Gerais*, *Federal University of Rio Grande do Sul* and *Federal University of Rio de Janeiro*) from the cohort of Top 401-500 Universities with similar PUB score and three universities (*Princeton University*, *California Institute of Technology* and *Swiss Federal Institute of Technology Lausanne*) from the Top 100 with the ARWU 2020 PUB score similar to previously mentioned seven universities.

Results

In total, 34,156 papers were scrutinised. As we can see from Table 1, *Princeton University*, *California Institute of Technology* and *Swiss Federal Institute of Technology Lausanne* lead the field with the percentage of papers published in Q1 journals. Although in official ARWU rankings, their score is similar to the seven remaining analysed universities for the PUB indicator. If we narrow down selecting SCIe and SSCI journals to Q1, the difference between universities is widened.

Table 1. Percentage of papers published in Q1-Q4 journals by the researchers from the observed universities

University	Q1	Q2	Q3	Q4
Belgrade	34.39	25.91	17.14	22.57
California	71.49	23.01	4.52	0.98
Lanzhou	50.92	28.28	12.40	8.40
Lausanne	65.85	24.32	7.55	2.28
Minas_Gerais	33.96	30.21	19.84	15.99
Nanjing	32.42	32.20	24.08	11.30
Princeton	68.87	22.30	6.68	2.15
Rio_de_Janeiro	37.39	30.87	19.58	12.16
Rio Grande_do_Sul	36.73	28.08	21.14	14.04
Xuzhou	40.20	26.28	21.80	11.73

Chinese universities (*Xuzhou* and *Lanzhou University*) exhibit excellent results that are aligned with the findings of Gao and Li (2020), which outlined that in 2018, the total number of papers published by Chinese colleges and universities in JCR Q1 journals exceeded 51,000 – an increase of about 18.6% over the year 2017. Brazilian universities are underachieving, with a considerable percentage of papers being published in Q4 journals that is aligned with the previous study (Mcmanus et al., 2020).

Furthermore, between the remaining seven universities, one can notice a significant difference ($p < 0.001$) in performance, with the *University of Belgrade* underperforming with 22.57% of the papers in lowest-ranked Q4 journals. If we dig a bit deeper (Table 2), we can witness many papers from the UB researchers being published in Serbian Q4 journals. Sadly, the trend continues with the similar findings presented in Zornic et al. (2014), Pilčević et al. (2018, 2019), with most of the papers from UB researchers being published in low-tier journals.

Table 2. Journals with the most considerable portion of papers published by the researchers from the University of Belgrade

Journal	Papers	%
VOJNOSANITETSKI PREGLED	94	2.7
SRPSKI ARHIV ZA CELOKUPNO LEKARSTVO	64	1.9
JOURNAL OF HIGH ENERGY PHYSICS	63	1.8
EUROPEAN PHYSICAL JOURNAL C	50	1.5
PHYSICS LETTERS B	44	1.3
ARCHIVES OF BIOLOGICAL SCIENCES	41	1.2
JOURNAL OF THE SERBIAN CHEMICAL SOCIETY	41	1.2
PHYSICAL REVIEW D	41	1.2
THERMAL SCIENCE	41	1.2

Conclusion

The potential inclusion of Q1 journals in ARWU methodology is not expected to trigger changes in the Top 100 tier. However, it will have profound consequences for lower-tier universities such as the University of Belgrade. Although the decline in ranking will be perceived negatively by the general audience, it can also serve as the catalyst of change and shift attention to researchers with admirable Q1 performance and consequently provide better support for their scientific endeavours. One of the potential directions for Universities' authorities might be including financial and other incentives for the publications in Q1 journals. Also, UB included SCIe and SSCI journals as a mandatory requirement for tenure back in 2008. Thus, one potential direction for UB' leaders might be strengthening the role of Q1 journals in academic promotion. In the future, this study could emphasise UB researchers who excel, which would contribute to the growing need to map excellence within the academic staff (Ioannidis et al., 2020; 2021).

References

ARWU (2021). ShanghaiRanking Consultancy, available at http://www.shanghairanking.com/

Gao, J., & Li, C. (2020). Version 2.0 of Building World-Class Universities in China: Initial Outcomes and Problems of the Double World-Class Project. *Higher Education Policy*, 1-17.

Ioannidis, J. P., Boyack, K. W. & Baas, J. (2020). Updated science-wide author databases of standardised citation indicators. *PLoS Biology,* 18(10), e3000918.

Ioannidis, J., Koutsioumpa, C., Vakka, A., Agoranos, G., Mantsiou, C., Drekolia, M. K., ... & Baas, J. (2021). Comprehensive mapping of local and diaspora scientists: a database and analysis of 63951 Greek scientists. bioRxiv.

Jeremić, V. & Stamenković, M. (2020). The Position of the University of Belgrade on the Shanghai List – Convergence Towards Reality, *Quarterly Monitor,* 62.

Mcmanus, C. M., Neves, A. A. B. & Maranhao, A. Q. (2020). Brazilian publication profiles: Where and how Brazilian authors publish. *Anais da Academia Brasileira de Ciências,* 92(2).

Mussard, M. & James, A. P. (2018). Engineering the global university rankings: gold standards, limitations and implications. *IEEE Access*, 6, 6765-6776.

Pilčević, I., Jeremić, V. & Vujošević, D. (2018). Evaluating the scientific performance of institutions within the university: An example from the University of Belgrade leading institutions. *Journal of the Serbian Chemical Society,* 83(11), 1285-1295.

Pilčevic, I., Bjeladinović, S. & Jeremić, V. (2019). The role of the integrated impact indicator (I3) in evaluating the institutions within a university. In ISSI 2019 proceedings (pp. 2706-2707).

Safón, V. (2019). Inter-ranking reputational effects: an analysis of the Academic Ranking of World Universities (ARWU) and the Times Higher Education World University Rankings (THE) reputational relationship. *Scientometrics,* 121(2), 897-915.

Sorz, J., Glänzel, W., Ulrych, U., Gumpenberger, C. & Gorraiz, J. (2020). Research strengths identified by esteem and bibliometric indicators: a case study at the University of Vienna. *Scientometrics,* 125(2), 1095-1116.

Zornic, N., Markovic, A. & Jeremic, V. (2014). How the top 500 ARWU can provide a misleading rank. *Journal of the Association for Information Science and Technology,* 65(6), 1303-1304.

Measuring Concentrations of Research Funding: A case of The Czech Science Foundation (GAČR)

Jiri Bures

bures@flu.cas.cz
Institute of Philosophy of the Czech Academy of Sciences, Jilská 1, 110 00 Praha (Czech Republic)

Introduction

In a comprehensive overview of current research on research funding, Aagaard et al. (2020) conclude that the subfield lacks a systemic approach to the analysis of funding concentrations. Independently, Katz and Matter (2020) applied established indices of wealth concentration (Gini coefficient, Palma ratio) to observe the distribution patterns of grants originating from *The National Institute of Health* (NIH). In this study, we respond to Aagaard et al. call for more empirical studies and precision in measuring funding concentrations by following the methodology established by Katz and Matter.

The Czech Science Foundation (GAČR) was founded in 1993 and is the largest public, multidisciplinary, basic research funding agency in Czechia. Since 2010, GAČR has covered on average 68 % (± 7.62) of all projects and 51 % (± 18.26) of all funds. Yearly, it provides on average 523 (± 78.1) projects with a success rate of 26.18 % (± 4.18) across all research areas.

Following this context, rationale, and subject, the study answers the question: What are levels of concentrations of funding originating from GAČR?

Data and methods

Data was collected from GAČR's *The Central register of R&D Projects* (CEP). The quality of the subset of projects was assessed using GAČR annual reports (Annual Reports 2021). The discrepancy between data and reports is -0.87 % (± 0.54) across years for projects and -0.83 % (± 1.79) for average size of funds per project. We assess discrepancy as being insignificant.

Following Katz and Matter (2020), concentrations were measured in "cumulative funding" (ibid p. 8), which is a sum of all funding that either principal investigator (PI-level) or organization (O-level) received in a year. Additionally, because GAČR is a multidisciplinary funding agency, grouping by disciplinary superset was applied (STEM, Social sciences, and humanities (SSH)).

After assessing correlations between Gini and Palma indices (figure 1), we have decided to use only the Palma index (following ibid p. 10). Being a ratio between bottom 40 % and top 10 % shares, it is more intuitive and sensitive to changes between bottom and top of population (Cobham et al., 2016).

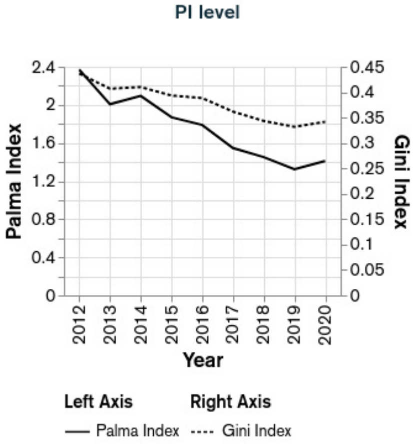


Figure 1. Correlation between Palma and Gini indices on PI-level across all disciplines.

Results

The concentration levels between STEM disciplines and SSH are far less significant than between organizational and PI's level of concentration (see fig. 2 and 3). O-level shows higher levels of concentration with stable development in observed time range, concentration on PI-level shows a decline. The difference between the Palma index is concretized for STEM in 2015 (see fig. 3). We observe that 10 % organizations accumulated 60 % of funding (13 PI), and top 10 % researchers were granted 33 % of funds (2.15 PI) .

Table 1. GAČR and NIH Palma ratio in 2015.

Agency	O-level	PI-level
GAČR*	13	2.15
NIH	~94/124**	~2.1/3.1**

**STEM projects **R and all funds (Katz & Matter, 2020: 8).*

Measured levels of concentrations are similar to those observed by Katz & Matter (2020) (see table 1). Specifically, difference between O and PI levels, as well as slow decline in funding concentration.

For NIH, from 3.5 to 3 for PI-level and from 150 to 124 for O-level.

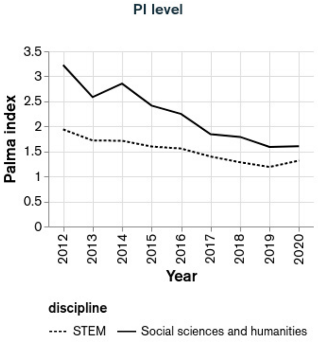


Figure 2. PI- level.

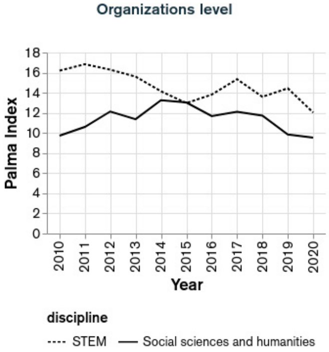


Figure 3. O-level.

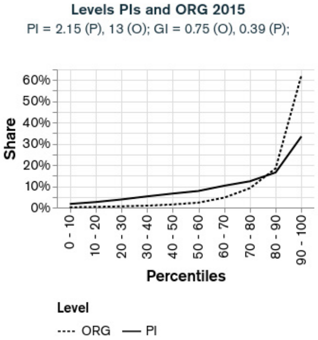


Figure 4. Accumulated share by percentiles of PI-level and O-level grantees in 2015.

Discussion

Two of our observations call for an immediate explanation. Firstly, we have observed a significant difference between concentration levels at O and PI levels. This may be due to the capacity of research organisations to hold not only more grants but also more "expensive" projects. Another reason may be that, unlike organisations, the amount of grants for individual scientists is capped by GAČR policies.

Secondly, we are seeing a similar concentration pattern in GAČR and NIH, especially the downward trend since 2010. For the NIH, this was explained as a response to public reflections about "research elites." A high difference for O-levels can be attributed to a different institutional structure and inaccuracy when comparing all STEM fields and specifically biomedicine (NIH).

Conclusion

This work is research in progress. We have measured levels of concentrations of GAČR's funding system and shown analytical potential of chosen research design. Future work will consist of extending the time dimension, including more agencies and leaving the STEM/SSH dichotomy for a more detailed disciplinary model. In the context of the research area, the study is the groundwork for a systematic approach in measuring and comparing various funding systems. Considering applied methodology, we are concluding that use of the Palma index offers a way to assess structure and change of funding concentrations.

Acknowledgments

This work was supported by the Czech Science Foundation, grant no. GJ20-01752Y, "Funded and Unfunded Research in the Czech Republic: Scientometric Analysis and Topic Modeling".

References

Aagaard, K., Kladakis, A. & Nielsen, M. W. (2020). Concentration or dispersal of research funding?. Quantitative Science Studies, 1(1), 117-149.

Annual Reports. (2021, January 22). The Czech Science Foundation (GACR). https://gacr.cz/en/annual-reports/

Cobham, A. , Schlögl, L. & Sumner, A. (2016). Inequality and the Tails: The Palma Proposition and Ratio, *Global Policy*, 7(1), 25-36.

Katz, Y. & Matter, U. (2020). Metrics of inequality: The concentration of resources in the US biomedical elite. *Science as Culture*, 29(4), 475-502.

Improving Coupling Metrics with Deep Neighborhood for Local Map of Science

Gaëlle Candel[1,2] and David Naccache[2]

{gaelle.candel, david.naccache}@ens.fr
[1] Worldline Labs, Paris (France)
[2] Département d'informatique de l'ENS, ENS, CNRS, PSL University, Paris (France)

Introduction

A citation graph is a directed acyclic graph where nodes are documents and links are reference relationships. For scientific papers, the citation graph can be used for multiple purposes: identify the most salient papers which receives a lot of attention from their community; analyze the collaborations between universities (Grauwin, 2011); or build a map of science to identify relationships between scientific areas (Boyack, 2005). The exploration of citation graphs was proposed by Garfield (2004), using similarity metrics between documents. The most used similarity metrics are Bibliographic coupling (BC) proposed by Small (1973) and Co-citation Coupling (CC) proposed by Kessler (1963). These metrics allow us to identify related papers locally. To obtain a larger overview, the generation of a map of science could help identify associated papers' communities.

Distribution over Citation Graph

BC and CC are very simple similarity measures, easy to put into practice. These metrics differ in multiple aspects. BC represents the authors' point of view, while CC represents the crowd's vision of the paper. At publication time, a paper is already coupled to other papers with BC, while uncoupled with any one with BC. Both evolve over time, leading to an increase in the number of related documents. As the number of references is fixed, the BC strength between two papers is fixed, while for CC, this strength can diminish or increase for each new citation received. Additionally, the number of references made by a paper is limited by authors' time and publication restrictions, such as the maximum number of pages allowed. Citations distribution follow a power-law distribution, while references distribution are less extreme. These asymmetries lead to a different outcome depending on which coupling metric is used. We follow the BC approach as it is more stable, and defined for a large majority of papers.

An issue with these types of metric is the graph sparsity. Several factors are leading to this property. *Independent discovery*: For two papers with a very similar idea, one new paper must choose to quote one or the other. *Admitted knowledge*: The community understands certain works so well that there is no longer any need to refer to them, such as principal component analysis. *Visibility*: search engines put the focus on top papers, reducing the visibility of newly published papers. Sparsity impacts reference overlap. For a document coupled with a thousand of documents, the overlap is rarely more than one. For two documents with 20 references each and 2 references in common, the Jaccard similarity score is close to 5%. Changing one coupling reference reduces it to 2.5%. This scoring with the first-order reference makes the score unstable to the reference omission. To limit this problem, documents can be aggregated by journal (Leydesdorff, 2009) or by the institution to artificially increase the number of references (Grauwin, 2011), or documents with insufficient coupling discarded, which drastically reduces the number of nodes and the granularity.

Deep Coupling

To improve coupling metrics, an idea is to look few step ahead to indirect references made by a paper's references. However, the items number grows as a^d, where $a$and $d$are the average number of references per papers and the exploration depth respectively. Metrics based on overlap only would give more credit to ancient knowledge which is not desirable. Instead, the contribution of distant older papers must be discounted. For this purpose, the indirect references $B_a$of $a$are weighted with a Random Walk with Restarts, give the weight $W_c^a$to a node $c \in B_a$. The similarity between two documents is obtained using the Cosine similarity:

$$S^{wc}(B_a, B_b, W^a, W^b) = \frac{\sum_{c \in B(a) \cap B(b)} (W_c^a W_c^b)}{\left(\sum_{c \in A} (W_c^a)^2 \sum_{c \in B} (W_c^b)^2\right)^{1/2}}$$

This measure allows obtaining a similarity score even for documents with no direct reference in common, as long as these documents belong to the same connected component of the graph.

Embedding with t-SNE

TSNE characteristics

TSNE (van der Maaten, 2008) is an embedding algorithm that has received a lot of attention from the machine learning community. This algorithm transforms a distance matrix into a $n \times d$vector, where two items close in the input space are close on the embedding. Items are displayed in a homogeneous scale, leading clusters' size to be proportional to the number of items they contain, allowing the identification of the main groups of interest.

Similarity to Distance

The Deep-BC allows defining the similarity between two nodes, with values in $[0,1]$. A distance metric has the inverse behaviour. The similarity score is simply transformed into distance using the equation $D^2 = (eps + S)^{-1} - (eps + 1)^{-1}$with the epsilon factor, set to $eps = 10^{-4}$ limits the maximal distances between completely unrelated items.

Dataset

We used the DBLP dataset (Tang, 2008), version 12, which contains 4,894,081 papers and 45,564,149 citation links. Documents are tagged, allowing an overview of the paper's topics and sample a subset of weakly related papers. Documents with no reference are discarded as our method cannot measure similarity on it.

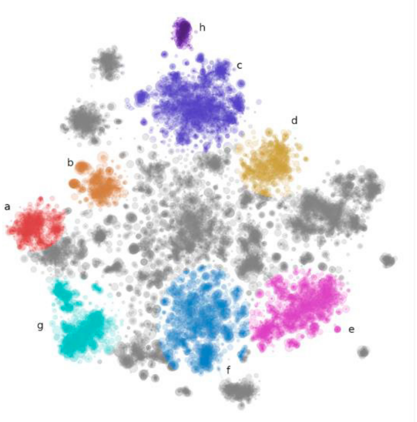


Figure 1. Embedding results for Biometrics. Some groups have been highlighted: a) *Keystroke analysis*, b) *Signature recognition*, c) *Security & Cryptography*, d) *Fingerprint recognition*, e) *Iris recognition*, f) *Facial recognition*, g) *Gait analysis*, h) *Smart card*.

Embedding results

Figure 1. shows the results for the Biometric topic. Dense groups are visually identifiable, each gathering papers from a sub-topic associated (*Iris rec.*) or distant to *Biometry* (Smart card, Security). Topics with similar technology (*Iris rec.*, *Facial rec.*, *Gait analysis* using Image processing; *Keystroke analysis*, *Fingerprint rec.* associated to behavioural analysis) are located nearby. This type of embedding can be easily clustered to extract communities automatically.

Conclusion

We proposed a method to embed a citation graph letting appears documents communities. Our approach is inspired by bibliographic coupling, extended with diffusion. The diffusion gives stability, which allows the preservation of documents with very few references. The embedding with tSNE allows obtaining well-spaced documents, helping to visually identify clusters, which could be used as a starting point for community analysis.

References

Boyack, K., Klavans, R. & Borner, K. (2005). Mapping the Backbone of Science. *Scientometrics*. 64. 351-374. 10.1007/s11192-005-0255-6.

Garfield, E. (2004). Historiographic Mapping of Knowledge Domains Literature. *Journal of Information Science*, *30*(2), 119-145. https://doi.org/10.1177/0165551504042802

Kessler, M. M. (1963), Bibliographic coupling between scientific papers. *Amer. Doc.,* 14, 10-25. https://doi.org/10.1002/asi.5090140103

Grauwin, S. Jensen, P. (2011). Mapping scientific institutions. *Scientometrics*, 89 (3), 943-954. 10.1007/s11192-011-0482-y. ⟨hal-00650267⟩

Leydesdorff, L. & Rafols, I. (2009). A Global Map of Science-Based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*. 60. 10.1002/asi.20967.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. & Su, Z. (2008). ArnetMiner: Extraction and Mining of Academic Social Networks. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (p./pp. 990--998), New York, NY, USA: ACM. ISBN: 978-1-60558-193-4

Small, H. (1973). Co-Citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science*. 24. 265-269. 10.1002/asi.4630240406.

van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*. 9, 2579-2605.

Gender of Editors in Chief and Editorial Member Committees in journals of Women Studies

Lourdes Castelló-Cogollos[1], Andrea Sixto-Costoya[2], Adolfo Alonso-Arroyo[3], Juan Carlos Valderrama-Zurián[4] and Rafael Aleixandre-Benavent[5]

[1] *lourdes.castello@uv.es*
Departament de Sociologia i Antropologia Social, University of Valencia. Spain. UISYS, Mixed Research Unit, CSIC-University of Valencia. Spain. Palau de Cerveró, Plaza Cisneros, 4, 46003 València (Spain)

{[2]andrea.sixto, [3] adolfo.alonso, [4] juan.valderrama}@uv.es
Departament d'Història de la Ciència i Documentació. Universitat de València, Spain. UISYS, Mixed Research Unit, CSIC-University of Valencia. Spain. Palau de Cerveró, Plaza Cisneros, 4, 46003 València (Spain)

[5] *rafael.aleixandre@uv.es*
UISYS, Mixed Research Unit, CSIC-University of Valencia. Ingenio (CSIC-Universitat Politècnica de València), Spain. Palau de Cerveró, Plaza Cisneros, 4, 46003 València (Spain)

Introduction

According to Merriam-Webster dictionary, women's studies is "the multidisciplinary study of the social status and societal contributions of women and the relationship between power and gender". The advancement of women in scientific and technological activities has been restricted by the structure of the editorial boards of journals in their respective areas. Running a scientific journal as Editor in Chief (EC) or belonging to its editorial board (EBM) is a recognition and honor that indicates advancement in one's career and allows to develop relationships with other colleagues, the formation of new professional networks and greater competence in obtaining resources (Capdeville et al., 2019; Macaluso et al., 2016). Based on previous experiences that reported that women are underrepresented on the EBM in the majority of scientific areas, the purpose of this paper was to examine this representation on the editorial boards in journals in a special scientific area in which women are the majority of contributors: the Subject Category Women Studies of Web of Science Core Collection (WS-WoS) database, analyzing their distribution by gender, countries and quartile. We assume that in this area, the representation of women in these committees should be higher.

Methods

The gender of the EC and EBM of the 45 journals included in 2019 in the subject category WS of the Social Science Citation Index (SSCI) of WoS was analyzed. For the compilation of each journal, its official pages or websites were consulted. For gender assignment of EC and EBM, the statistical package Genderize.io (https://genderize.io) was used, which provides a probability of male or female gender based on a frequently updated database that currently includes greater than 200,000 distinct first names from more than 79 countries and languages. From the data collected, the following indicators were calculated: gender distribution of EC and EBM by journal, publishers, quartile of the journal in JCR and country of affiliation of members. In addition, data on the countries of membership were grouped into three major blocks: United States, European Union and rest of the World.

Results

The total number of editorial members was 1,362. Among the EBM, 1,119 were female (82.16%) and 243 (17.84%) male, and among EC, 69 (92%) were female and six (8%) were male. In five of the 45 journals, all EBM were women. Only two men from the six EC manage a journal as the sole EC, and in the other four journals, a man shared the EC with other women. The EBM and EC belonged to 65 countries. In 28 countries all EBM were women and in 10 countries all EC were women; only in 9 countries were all men. Among the countries with more than 10 members, the highest percentage of women corresponds to France (92.59%), Australia (89.06%) and South Africa (88.24%), and the lowest to South Korea (56%), Brazil (58.33%) and New Zealand (61.54%) (figure 1). The highest percentage of women corresponds to Europe (83.62%); in the United States it is 82.88% and in the rest of the world it is 79.12% (figure 2). In the gender distribution by quartiles of journals in JCR, it is observed that the highest percentage of women occurs in journals belonging to second quartile (84.29%) and the lowest in journals belonging to fourth quartile (79.29%) (table 1).

Figure 1. Percentage of women and men of EC and EBM by countries.

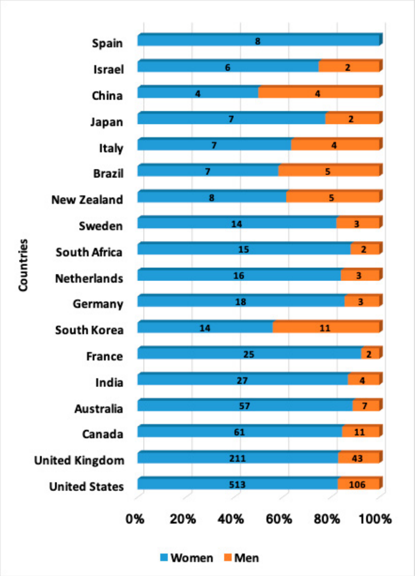


Figure 2. Number and percentage of women and men of EC and EBM by large geographic areas.

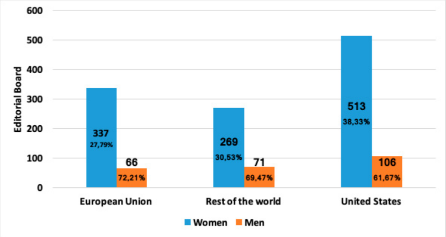


Table 1. Gender distribution of EBM and EC by quartile of journals in JCR.

		EBM and EC				
Cuartile	Nº Journals	Women	% women	Men	% men	Total
Q1	11	284	80.68	68	19.32	352
Q2	11	381	84.29	71	15.71	452
Q3	11	232	83.45	46	16.55	278
Q4	12	222	79.29	58	20.71	280
Total		1,119		243		1,362

Discussion

Previous studies reported a low presence of women on EBM and EC of scientific journals in most areas that does not reflect the gender composition in the corresponding specialties. Is the case, for example, of Medicine (from 1.6% to 41.7%) (Alonso-Arroyo et al., 2021), except for specialties in which women have historically predominated as Nursing, Physical Therapy and other related to Nutrition (lactation and breastfeeding), and General Marketing and Management (22%) (Pan & Zhang, 2013; Metz & Harzing, 2012). However, in the Women Studies area of the JCR, the situation is quite the opposite, since women are better represented in the majority of the journals, in some cases accounting for the total number of editorial members. The high representation of women in the area of WS offers a counterpoint to the structural and cultural conditioning factors which leads to women being under-represented in most of areas, as well as a powerful answer to stereotypical assessments of women's lower qualifications for leadership positions. The reasoning for many of the arguments proposed for the low participation of women in most subject areas is, therefore, insufficient, and the results in the area of WS are proof of this. It confirms that achieving gender equality is, at least in the field of professional and scientific publication, an attainable and solvable objective. Future work could analyze the gender composition of reviewers as a measure of potential future gender diversity, due that potential members of editorial committees are commonly selected from among reviewers.

References

Alonso-Arroyo, A., González de Dios, J., Aleixandre-Agulló, J. & Aleixandre-Benavent, R. (2020). Gender inequalities on editorial boards of indexed pediatrics journals. *Pediatric Research*. doi: 10.1038/s41390-020-01286-5 .

Capdeville, M. (2019). Don't Hold Your Breath-The Rise of Women on Journal Editorial Boards. J *Cardiothoracic Vascular Anesthesia*, 33, 3235-3238.

Macaluso, B., Larivière, V., Sugimoto, T. & Sugimoto, C. R. (2016). Is Science Built on the Shoulders of Women? A study of gender differences in contributorship. *Academic Medicine*, 91, 1136-42.

Metz, I. & Harzing, A. (2012). An update of gender diversity in editorial boards: a longitudinal study of management journals. *Personnel Review*, 41, 283-300.

Pan, Y. & Zhang, J. Q. (2014). The composition of the editorial boards of General Marketing journals. *Journal of Marketing Education*, 36, 33-44.

Differences between using institutional information from curriculum vitaes and publications for tracking scientific mobility

Yu-Wei Chang

yuweichang2013@ntu.edu.tw
Department of Library and Information Science, National Taiwan University, Taipei (Taiwan)

Introduction

According to a review article by Gureyev et al. (2020) and a literature search of the Scopus database at the end of 2020, research on the mobility of researchers has been increasing since 2017. In particular, several studies have employed the bibliometric method—the analysis of publications' bibliographic records including author affiliation information—to track the academic trajectory of researchers in terms of which institutions they were affiliated to during a certain period (Robinson-Garcia et al., 2019; Sachini, Karampekios, & Brutti, 2020). In addition to the bibliometric method, analyzing curriculum vitaes (CVs), which provide a summary of a researcher's academic background and experience, is another nonintrusive method to identify a researcher's institutional footprint (Gureyev et al., 2020).

The bibliometric method requires less effort for collecting and processing data than does CV analysis because a single literature database can provide the required sample data; however, the data revealed in researchers' publications and CVs are distinct. CVs reveal the purpose of each institutional move, such as earning a PhD degree or serving as a visiting scholar, whereas publications reveal only the year of the publication and author affiliation. The situation of researchers with low productivity or publications that are not indexed by literature databases leads to a loss of certain institutional mobility data. The CV method also results in some difficulties during research on scientific mobility because of incomplete personal information. The limitations of CV and bibliometric analyses make the combination of the two methods desirable.

Although a few studies have combined CV and publication data (e.g., Fangmeng, 2016; Jonkers & Tijssen, 2008), these studies have not focused on the different types of institutions that researchers were affiliated to and that they visited during their academic career. To address this research gap, the following main research question is examined in this study: Does the use of CV analysis and publication analysis result in significantly different findings regarding the number of institutions that researchers were affiliated to and visited? The current results may reveal the appropriate method for exploring scientific mobility.

Methodology

A total of 416 recipients of the Sloan Research Fellowship in Mathematics between 1990 and 2019 were selected as the sample researchers. This fellowship is awarded to distinguished junior mathematics researchers. Therefore, the recipients were assumed to have remarkable research performance, resulting in a higher possibility that they may transfer to institutions with greater research resources and a higher reputation than their previous institution.

The names of the recipients and their institutional affiliations when they obtained the fellowship were the basic information used for searching their CVs and other biographical information on the Internet. The institutions where these researchers received their bachelor's, master's, and doctoral degrees were distinguished from those where these researchers were affiliated to or visited temporarily. The period at each institution was also recorded, if available, for the researchers.

Each recipient's author affiliation information and the publication date listed in their publications, obtained from the author profile on Scopus, were retrieved as another information source for monitoring scientific mobility. The CV data helped correctly identify the recipients' author profile on Scopus.

The institutions that the sample researchers were affiliated to after receiving their PhD degrees were divided into home and temporary institutions according to CV data. Home institution was defined as the researchers' affiliated institution, and temporary institution was defined as any institution that the researchers visited temporarily for periods not exceeding 1 year. Institutions not included in CVs and those at which the researcher published their research for more than 1 year were categorized as home institutions, and other institutions were regarded as temporary ones. The main focus of this study was to compare the number of institutions yielded using the CV and bibliometric methods.

Results

Most researchers had visited other institutions for periods ≤1 year according to both their CVs (66.6%) and publications (62.5%; Table 1). The average numbers of home institutions, temporary institutions, and all institutions per researcher according to CV analysis were higher than those yielded from bibliometric analysis. This indicates that some researchers did not publish with the intention of giving credit to certain home and temporary institutions. Moreover, both CV and publication analyses revealed that the average number of temporary institutions per researcher was higher than that of home institutions. CV data indicated that the average total number of institutions per researcher was 4.42, which is higher than the number (3.96) identified using the publication method. A weak positive correlation was noted in the total number of home and temporary institutions between CVs and publications (correlation coefficient: 0.344).

Table 1. Institutional information by information source

	CVs	Publications
Researchers (H)	139(33.4%)	156(37.5%)
Researchers (H&T)	277(66.6%)	260(62.5%)
Ave N. (H) / (T)	2.50 / 5.38	2.42 / 4.88
Ave. N. (H&T)	4.42	3.96

Note: H, home institution; T, temporary institution; H&T, home and temporary institutions.

Half of the researchers (50%) with home and temporary institutions listed on their CVs also had records of the two categories of institutions in their publications (Table 2). This means that consistent institutional data could be obtained from the two data sources. For 17% of the researchers who had visited other institutions, that data were not found in their publications. A surprising finding is that 12% of the researchers had no records of visiting other institutions in their CVs.

Table 2. Number of recipients by category of institution according to CV and publication data

Group	Number (%)
CV(H&T) & P(H&T)	208(50)
CV(H) & P(H)	87(21)
CV(H&T) & P(H)	69(17)
CV(H) & P (H&T)	52(12)

Note: P, publication; CV, curriculum vitae.

Table 3 lists 14 categories of institutional information that integrate both the numbers and names of institutions between CVs and publications. Only 26% of the researchers in Group B had an equal number and the same home and temporary institutions in both CVs and publications. These researchers differed from those in Group E, who had an equal number but not the same institutions. This indicates that the information of institutional mobility of most researcher was inconsistent between CVs and publications.

Table 3. Comparison of numbers and categories of institutions according to CV and publication data

Category		N(%)
A	CV(H) = P(H)* & CV(T) > P(T)	116(27.9)
B	CV(H&T) = P(H&T)*	108(26.0)
C	CV(H) = P(H)* & CV(T) < P(T)	67(16.1)
D	CV(H) > P(H) & CV(T) > P(T)	27(6.5)
E	CV(H&T) = P(H&T)	23(5.5)
F	CV(H) < P(H) & CV(T) < P(T)	19(4.6)
G	CV(H) > P(H) & CV(T) < P(T)	16(3.8)
H	CV(H) > P(H) & CV(T) = P(T)*	15(3.6)
I	CV(H) < P(H) & CV(T) = P(T)*	10(2.4)
J	CV(H) < P(H) & CV(T) > P(T)	8(1.9)
K	CV(H) > P(H) & CV(T) = P(T)	3(0.7)
L	CV(H) = P(H) & CV(T) > P(T)	2(0.5)
M	CV(H) = P(H) & CV(T) < P(T)	1(0.2)
N	CV(H) < P(H) & CV(T) = P(T)	1(0.2)

Note: * refers to an equal number and the same institutions.

Conclusion

This study confirmed that neither CV analysis nor bibliometric analysis provided complete information regarding the institutions that most researchers were affiliated to and visited. Combining the two methods can improve the precision of results, although more time and financial resources are necessary. To obtain a more complete understanding of the two methods, future research should compare them over different periods.

Acknowledgments

This research was supported by a grant from the Ministry of Science and Technology of Taiwan (MOST 109-2410-H-002-207).

References

Gureyev, V. N., Mazov, N. A., Kosyakov, D.V., & Guskov, A. E. (2020). Review and analysis of publications on scientific mobility: Assessment of influence, motivation, and trends. *Scientometrics, 124*(2), 1599–1630.

Jonkers, K., & Tijssen, R. (2008). Chinese researchers returning home: Impacts of international mobility on research collaboration and scientific productivity. *Scientometrics, 77*(2), 309–333.

Robinson-Garcia, N., Sugimoto, C. R., Murray, D., Yegros-Yegros, A., Larivière, V. D, & Costas, R., (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics, 13*(1), 50–63.

Sachini, E., Karampekios, N., Brutti, P., & Sioumalas-Christodoulou, K. (2020). Should I stay or should I go? Using bibliometrics to identify the international mobility of highly educated Greek manpower. *Scientometrics, 125*(1), 641–663.

Tian, F. (2016). Brain circulation, diaspora and scientific progress: A study of the international migration of Chinese scientists, 1998-2006. *Asian and Pacific Migration Journal, 25*, 296–319.

Mapping the Interdisciplinarity of the Arts, Humanities & Social Sciences by Paper-level Classifications of Science

Bikun Chen[1], Shiqin Wang[2], Yuxin Liu[3] and Mengxia Cheng[4]

[1] chenbikun@njust.edu.cn, {[2] 939226312, [3] 283958279, [4] 1291361237}@qq.com
Nanjing University of Science and Technology, Nanjing (China)

Introduction

Most interdisciplinary researches in scientometrics field rely on journal-level classifications of science. By this approach, papers could be misclassified in journal classification systems (Shu et al. 2019), which may cause bias, especially if there is a significant proportion of multidisciplinary journals in the reference list. In order to avoid these limitations, direct citation, co-citation, bibliographic coupling between publications are applied (eg. Waltman & Eck, 2012).

Different from international journal-level classification system (eg. WoS and Scopus Categories), most Chinese bibliographic databases (eg. CNKI and Wanfang Data) classify publications at the paper level using the Chinese Library Classification Scheme (CLC) (Chinese Library Classification 2010). CLC is also used by all the publishers in China to classify all publications including journals, books, and monographs. Based on paper-level classifications of science, 55,894 research articles published from 2008 to 2018 in 20 core Chinese journals that belong to "Library, information and archival science" are used to map interdisciplinary network of tier-1, 2, 3 and 4 classifications of science (Chen et al., 2019).

As an extension of the former research (Chen et al., 2019), all articles of the Arts, Humanities & Social Sciences published from 2008 to 2018 are incorporated to map interdisciplinary network and more interdisciplinary measures are calculated to measure interdisciplinarity in this study.

Data

The raw data consist of 1,153,838 articles (only 991,034 academic articles are kept in this study) published from 2008 to 2018 in 568 core Chinese journals indexed by CSSCI (Chinese Social Sciences Citation Index, 2019-2020 Version) (http://cssrac.nju.edu.cn/index.html). The raw data are crawled from March to May, 2020 in CNKI (China National Knowledge Infrastructure) (https://oversea.cnki.net/index/) and manipulated from June to August, 2020. Every Chinese journal has a tier-1 or 2 CLC code (detailed in Table 1) and almost each Chinese article has at least one tier-2, 3, 4, 5 or 6 CLC code (detailed in Table 2). Besides, CSSCI has its own unique journal classification system (not strictly follow CLC).

Table 1. Sample journal statistics.

CLC Code	# of Journals	# of Papers
Tier-1	447 (78.7%)	720,482 (72.7%)
Tier-2	121 (21.3%)	270,552 (27.3%)

Table 2. Sample paper statistics.

CLC Code	# of Papers	# of Papers (CLC Code >= 2)
Tier-1	991,034 (100%)	200,464 (20.2%)
Tier-2	991,021 (99.9%)	200,464 (20.2%)
Tier-3	978,341 (98.7%)	197,974 (20.2%)
Tier-4	896,381 (90.5%)	187,059 (20.9%)

Method

Pierce (1999) summarized three ways to conduct interdisciplinary research: borrowing, collaboration and boundary crossing. Based on the theories, two methods are used to construct interdisciplinary networks and network coherence indicators are applied to measure interdisciplinary. Besides, interdisciplinarity diversity of tier-1 CSSCI categories are calculated.

Method 1 (Undirected network): Co-occurrence of article-level CLC codes (>= 2) (proxy of borrowing). When any article-level CLC code i and j co-occur in any article A, their co-occurring value is number one. The total relation intensity φ between CLC code i and j is the sum of CLC code i and j co-occurring in any article A. The total relation intensity is normalized by Jaccard Index.

$$\emptyset_{i\,j} = \sum_{A} One(CLC_i, CLC_j)$$

Method 2 (Directed network): Article-level CLC codes point to journal-level CLC code (proxy of publishing work in other disciplines). When any article-level CLC code k points to journal-level CLC code l in any article B, their value is number one. The total relation intensity φ between CLC code k and l is the sum of CLC code k points to l in any article B. The total relation intensity is also normalized.

$$\emptyset_{k\,l} = \sum_{B} One(CLC_k, CLC_l)$$

Interdisciplinary Measures (Diversity & Network coherence): Different interdisciplinarity diversity

and network indicators are calculated, such as number of specialties, Specialization index, Brillouin index, network density, diameter, average node degree, average path length, average cluster coefficient, node centrality and so on.
Besides, tier-4 categories can be upgraded to tier-3, 2 and 1 categories. So, all upgraded categories are also included in each upper-level network.

Results

Due to the limited pages, only the results of tier-1 categories are shown in Table 3, Figure 1 and 2. From Table 3, "Ethnology & Culturology", "University Journal", and "General Social Science" keep a balance between variety and evenness. From Figure 1 and 2, each node indicates a category (labeled by CLC code), node size indicates its weighted degree, node colour indicates its category type (blue nodes indicate Arts & Humanities, green nodes indicate Social Sciences, red nodes indicate Natural Science). In Figure 1, "Economics" (F), "Culture, Science, Education & Sports" (G) and "Politics & Law" (D) are most closely related. In Figure 2, "General Social Science" (C), "Economics" (F), "Culture, Science, Education & Sports" (G) and "Politics & Law" (D) are most boundary- crossing.

Table 3. Interdisciplinarity diversity of tier-1 CSSCI categories (top 10 Brillouin Index).

Tier-1 CSSCI Category	Brillouin Index	# of Specialties	Specialization Index
Ethnology & Culturology	0.980	21	0.119
University Journal	0.960	21	0.138
General Social Science	0.913	22	0.157
Natural Resources & Environment	0.744	19	0.246
Philosophy	0.698	21	0.249
Human& Economic Geography	0.704	20	0.282
Marxism Theory	0.716	21	0.310
Sociology	0.721	20	0.359
History	0.609	22	0.406
Politics	0.655	22	0.406

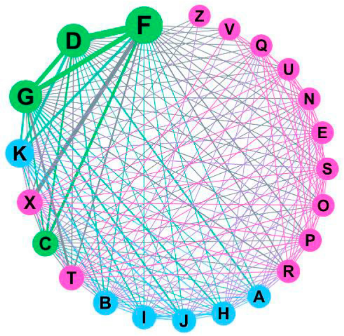


Figure 1. Undirected network of tier-1 categories.

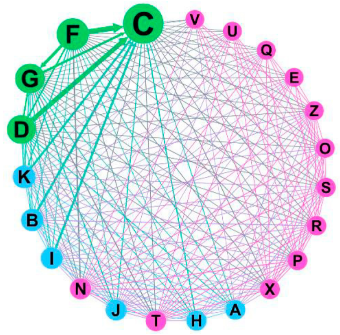


Figure 2. Directed network of tier-1 categories.

Acknowledgments

This paper is supported by Key Project of National Social Science Fund in China (16ZDA224).

References

Chen B, Cheng M, Li P, et al. (2019). Interdisciplinary research based on paper-level classifications of science- A preliminary case study of Chinese journals. In Catalano G., et al. (Ed.), *Proceedings of the 17th International Conference on Scientometrics and Informetrics* (ISSI 2019) (pp. 2718-2719). Rome: ISSI.

Pierce, S. J. (1999). Boundary crossing in research literatures as a means of interdisciplinary information transfer. *Journal of the American Society for Information Science*, 50(3), 271-279.

Shu, F., Julien, C., Zhang, L., Qiu, J., Zhang, J. & Lariviere, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics*, 13(1), 202-225.

Waltman, L. & Eck, N. J. V. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378-2392.

Zhongguo Tushuguan Fenleifa [Chinese Library Classification] (2010). (5 ed.). Beijing: National Library of China Publishing House.

WeChat Presence of Chinese Scholarly Journals: An analysis of CSSCI-indexed Journals

Ting Cong[1], Zhichao Fang[2] and Rodrigo Costas[3]

[1] *congting13@163.com*
University of Shanghai for Science & Technology, Dept of Publishing, Jungong Road 516, Shanghai City (China)

{[2] z.fang, [3] rcostas}@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (Netherland)

Introduction

How science is discussed on social media has become an increasingly popular topic in scientometrics, with most studies focusing on global platforms such as Twitter, Facebook, and ResearchGate. However, the online discussion of science on *local* platforms has been seldom studied, particularly due to the barriers of languages and data infrastructures for analyzing “local altmetrics” (Yu et al, 2017). China, as a non-English speaking country with the largest number of researchers and internet users, has some indigenous social media platforms mostly used by Chinese users, such as Sina Weibo and Tencent WeChat. These social media provide institutions and individuals with platforms not only for entertainment but also for scholarly communication. In China, WeChat is one of the most popular social media platforms for academics to publicly communicate about research developments. According to Zhang’s research (2018), 53.1% of scholarly journals indexed by Chinese Social Sciences Citation Index (CSSCI) had established WeChat accounts until 2017. This study also aims at investigating the WeChat presence of CSSCI-indexed journals. Specifically, the following research questions are addressed:

RQ1. How many CSSCI-indexed journals have WeChat public accounts?

RQ2. At the journal level, how is the correlation between bibliometric indicators and WeChat indicators?

RQ3. Do journals with or without WeChat public accounts perform differently in terms of bibliometric indicators?

Data and Methods

Data collection

CSSCI stands for Chinese Social Sciences Citation Index (http://cssci.nju.edu.cn/), a selective citation index which covers Chinese scholarly journals in the field of Social Sciences and Humanities. In the 2019-2020 edition of CSSCI, there are a total of 782 journals indexed in its core and expanded collections (including 568 journals from the core and 214 from the expanded collection). For each of these journals, we harvested a range of bibliometric data from the database of China National Knowledge Infrastructure (CNKI) between September 28, 2020 and October 8, 2020, such as total number of published papers, total number of citations, total number of downloads, and journal impact factors.

To examine whether these journals have created WeChat public accounts, we manually searched the journals’ names in the built-in search engine on WeChat. For those journals with WeChat public accounts with posts, we used a third-party software to collect the detailed information for each of their WeChat posts (e.g., title, number of clicks, number of likes). The WeChat data collection was conducted during the period from October 10, 2020 to November 9, 2020.

Analytic approach

Figure 1 illustrates the research workflow of the study. Starting from the extraction of both WeChat and bibliometric data for the CSSCI-indexed journals, followed by their aggregation at the journal-level, and finally performing three types of analysis: general descriptive statistics, correlation and factor analysis and finally a comparative analysis of citation performance between journals with and without WeChat accounts.

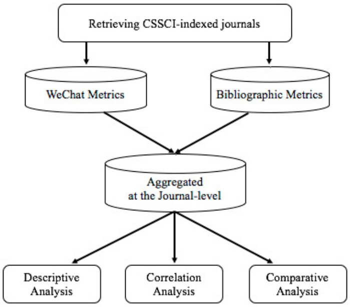


Figure 1. Research workflow

Results

Presence of CSSCI-indexed journals on WeChat

Among the 782 CSSCI-indexed journals, 511 of them have created WeChat public accounts (accounting for 65.3%). However, there are 6 journals with WeChat public accounts that never posted any WeChat post until the dater of data collection. Therefore, we collected WeChat activity indicators for a total of 505 journals. The total number of WeChat posts by CSSCI journals is 193,367. These posts have received a total of 272.9 million clicks and 1.2 million likes, respectively. These large numbers reflect the active interactions that take place in WeChat.

Correlation analysis of bibliometric and WeChat indicators

In this section, we focus on the correlation analysis of bibliometric and WeChat indicators. To examine the relationships between WeChat indicators and bibliometric indicators, Spearman correlation analyses were conducted as shown in Figure 2. Within both bibliometric indicators and WeChat indicators, indicators are moderately or strongly correlated. However, the correlations between bibliometric and WeChat indicators are generally weak or negligible. It suggests that bibliometric and WeChat social media indicators capture substantially different types of interactions and impact.

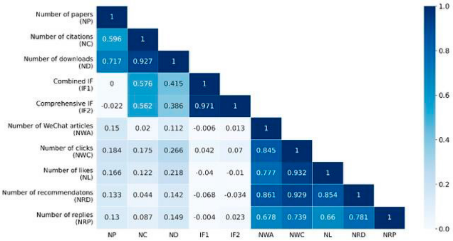


Figure 2. Spearman correlation analyses among bibliometric and WeChat indicators

Comparison of bibliometric indicators between journals with and without WeChat accounts

In this section, we explore whether journals with WeChat accounts have higher values of bibliometric indicators. Figure 3 plots the performance of journals with and without WeChat public accounts from the perspectives of the number of published papers, citations, downloads, and the two journal impact factors. In general, journals with WeChat public accounts tend to have slightly higher bibliometric scores than those without WeChat public accounts.

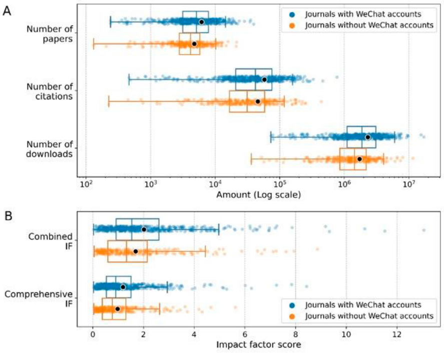


Figure 3. Bibliometric indicators of journals with and without WeChat public accounts

Discussion and Conclusion

This study empirically studied the presence of CSSCI-indexed journals on the Chinese WeChat platform, falling into the category of altmetric studies with a focus on "local research discussed in local platforms" (Yu et al., 2017). The findings of this research contribute to a better understanding of how Chinese scientific journals use local social media platforms to reach broader audiences. Although still in progress, this research already hints to important conclusions that highlight the relevance of incorporating local perspectives in the study of social media metrics across countries.

References

Costas, R., Zahedi, Z. & Wouters, P. (2015). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10), 2003-2019.

Han, Z., Htet, H., Mojisola, E., Tai-Quan, P., Aravind, S. & Yin-Leng, T. (2019). Social media presence of scholarly journals. *Journal of the Association for Information Science and Technology*, 70(3), 256-270.

Xiaoqiang, Z., Yuan, J. & Bin, Y. (2018). Operation on WeChat official accounts of Chinese academic journals with high WCI and its implications. *Chinese Journal of Scientific and Technical Periodicals*, 29(6), 574-584. (***in Chinese***)

Yu, H., Xu, S., Xiao, T., Hemminger, B. M. & Yang, S. (2017). Global science discussed in local altmetrics: Weibo and its comparison with Twitter. *Journal of Informetrics*, 11(2), 466-482. https://doi.org/10.1016/j.joi.2017.02.011

Understanding Scientific Success: A MacIntyrean Virtue Ethics Approach

Cinzia Daraio

daraio@diag.uniroma1.it
Sapienza University of Rome, Department of Computer Control and Management Engineering A. Ruberti (DIAG), via Ariosto, 25 00185 Rome (Italy)

Introduction

The analysis of the success in science has been a subject of interest and study for a long time. "Success breeds success" or the so called Mattheu effect, as it was named by Merton (1973), was used to explain the cumulative effects observed in the sociology of science, analysing noble price winners. But what are the determinants of scientific success? What makes a tiny proportion of scholars successful? Why a few scholars rise far above the rest, or using the terms of de Solla Price (1963, p. 59), why there are a "few giants and a mass of pygmies" and "neither man nor nature pushes us toward egalitarian uniformity"?

In this work we propose the adoption of a MacIntyrean virtue ethics approach to understand scientific success.

Related Literature

Scientific talent, according to Feist (2013), is nature shaped by nurture. In the "Complexity of Greatness" Kaufman (2013) combines different views and put together research on genes, talents, intelligence and expertise, deliberate practice, creativity prodigies, savants, mindset, passion and persistence to show that success is much more complex than talent and practice. Pluchino et al. (2018) develop an agent-based model which shows that talent is necessary to be successful in life, but the most talented people almost never reach the highest peaks of success, being surpassed by averagely talented but sensibly luckier individuals. Often in the literature success is connected to luck (Pritchard, 2005; Mauboussin, 2012; Frank, 2016). Barabási (2018) defines success as "the rewards we earn from the communities we belong to" and investigates the networking skills that enhance success. Barabási, distinguishing performance from success, states that personal achievements are important but in order to be translated into success, performance requires a community reaction. Hence, success is a collective undertaking. Besides performance, success requires building trust and reputation in order to shape others' perceptions regarding one's achievements. Barabási (2018) illustrates the network science behind why people succeed or fail presenting the following five universal laws of success: *i)* Performance drives success, but when performance can't be measured, networks drive success; *ii)* Performance is bounded but success is unbounded; *iii)* Previous success x fitness = future success; *iv)* While team success requires diversity and balance, a single individual will receive credit for the group's achievements, and *v)* with persistence success can come at any time.

These five rules state that a scientist place in a network is based not only on her performance, but also on the network's perception of her contribution to its goals. Day (2019) in reviewing Barabási (2018) identifies a weakness of the book in "the distance between the simplicity of how Barabási communicates the so-called laws of networks and the actual ability required to identify and master networks in everyday life". We believe that adopting a MacIntyrean approach to understand success can help to fill in this existing gap.

Aim and contribution

The aim of this work is to show that the adoption of the modern ethics of virtues developed by the moral philosopher Alasdair MacIntyre can offer us an interesting broader framework for getting into the mechanisms of scientific success. Using philosophical argumentation, we try to show how the MacIntyrean Aristotelian-Thomistic ethics of virtues allows us to understand a little better what lies behind scientific success by considering *virtues* that are stable dispositions of the character, connected to the psychology and motivation of individuals.

Materials and Method

MacIntyre focuses insistently on the connection between ethical doctrines and the historical-social contexts in which they arise, advancing arguments in favor of a "community" ethics which, unlike the abstract models developed by utilitarianism and neo-contractualism, identifies the ethically relevant principles in the *concrete community* to which individuals belong. In this perspective, his works are aimed at identifying historically the most satisfactory ethical models, and among these he attributes particular importance to the Aristotelian theory of virtues and to the Thomist tradition. Our proposal is based on the consideration of research as a social practice (MacIntyre, 2007). MacIntyre's concept of social practice seems particular suitable for conceptualizing the communal nature of research as illustrated in Bezuidenhout (2017), where scientists mediate their social interactions with peers during daily laboratory research. In Daraio and Vaccari (2020, 2021) applying this notion of research we have identified leaders, good researchers and honest researchers who contribute with their duties and activities to the achievement of the "internal goods" of research practices aiming at overcoming existing knowledge. Making a step further, MacIntyre (1999) develops the concept of "networks of giving and receiving" that are grounded on relationships that involve affective and sympathetic links which are grounded in and governed by norms of uncalculated and unpredicted giving and receiving and the concept of "virtues of acknowledged dependence", the most important of which is *misericordia* or mercy. These relational virtues seem closely connected to the social nature of research practices.

Towards the assessment of success: the contribution of virtue ethics

Extending the evaluation framework to include the ethics of virtues allows us to take into consideration the *character* (Miller et al., 2015) of the researchers and to understand and value how the stable motivations and the traits of the character (i.e. virtues) contribute to the research practices. Snow (2010) shows that virtue may be considered as social intelligence and Albrecht (2006) on a different ground identifies social intelligence as the new science of success. Hence, our framework offers us the possibility to *i)* understand the functioning of research practices in which there is a division of labour, including different kind of researchers and *ii)* design monitoring and controlling systems well suited for the specificities of the activities carried out within research practices.

If we deepen our understanding on why there are "giants" and pygmies within research practices, we can have a better understanding on what is research and what does it mean to carry out a good research. Further this understanding can be helpful to design "good evaluations of scientific research" (those that give values and promote good research) and to discriminate between good and bad research practices. Clearly success has an interplay with luck and merit (Daraio, 2021).

Concluding remarks and further research

The connection between performance, success, merit and luck still remains to be further investigated from an ontological and epistemological point of view. We try to show that an Aristotelian-Thomist moral framework is useful to clarify some connections among these relevant concepts.

Acknowledgments

The funding support of Sapienza Awards n. PH11715C8239C105 and n. RM11916B8853C925 is gratefully acknowledged. This work is prepared within the Joint Diploma St Thomas Aquinas Doctor Humanitatis 2020-2021.

References

Albrecht, K. (2006). *Social intelligence: The new science of success*. John Wiley & Sons.

Barabási A. L. (2018). *The Formula: The Universal Laws of Success, The Science Behind Why People Succeed or Fail*, Hachette Book Group, New York.

Bezuidenhout, L. (2017). The relational responsibilities of scientists:(Re) considering science as a practice. *Research Ethics*, 13(2), 65-83.

Day, M. (2019). Albert-Lászlo Barabási, The Formula: The Universal Laws of Success. The Science Behind Why People Succeed or Fail. Book Review, *International Journal of Communication*, 13, 4, 5595-5598.

Daraio, C. & Vaccari A. (2021). Perché è importante fare una buona valutazione della ricerca. La proposta delle virtù, *Bollettino della Società Filosofica Italiana*, gennaio-aprile 2021, 45-59.

Daraio, C. (2021). In Defense of Merit to Overcome Merit*, Frontiers in Research Metrics and Analytics*, January 2021, Volume 5, Article 614016. doi: 10.3389/frma.2020.614016.

De Solla Price, D. J. (1963). *Little science, big science... and beyond*. New York: Columbia University Press.

Feist, G. (2013). Scientific talent: Nature shaped by nurture, in Kaufman eds., *The complexity of greatness: Beyond talent or practice*, Oxford University Press, 257-274.

Frank, R. H. (2016). *Success and Luck: Good Fortune and the Myth of Meritocracy,* Princeton University Press.

Kaufman, S. B. (Ed.). (2013). *The complexity of greatness: Beyond talent or practice*. Oxford University Press.

MacIntyre, A. (2007). *After Virtue: A study in Moral Theory*, 3rd Ed. Notre Dame: University of Notre Dame Press.

MacIntyre, A. (1999). *Dependent rational animals – why human beings need the virtues*. Carus Publishing Company Illinois.

Mauboussin, M. J. (2012). *The success equation: Untangling skill and luck in business, sports, and investing.* Harvard Business Review Press.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations.* University of Chicago press.

Miller, C. B., Furr, R. M., Knobel, A. & Fleeson, W. (Eds.). (2015). *Character: New directions from philosophy, psychology, and theology*. Oxford University Press.

Pluchino, A., Biondo, A. E., & Rapisarda, A. (2018). Talent versus luck: The role of randomness in success and failure. *Advances in Complex systems*, 21(03n04), 1850014.

Pritchard, D. (2005). *Epistemic luck*. Oxford Un Press.

Snow, N. E. (2010). *Virtue as social intelligence: An empirically grounded theory*. Routledge.

Effect of Firm Size on Patent Maintenance Length

Huei-Ru Dong [1,3] and Mu-Hsuan Huang[2,3]

[1] *141646@mail.fju.edu.tw,* [2] *mhhuang@ntu.edu.tw*

[1] Fu Jen Catholic University, Dept of Library and Information Science, New Taipei City (Taiwan)

[2] National Taiwan University, Dept of Library and Information Science, Taipei (Taiwan)

[3] National Taiwan University, Center for Research in Econometric Theory and Applications (CRETA), Taipei (Taiwan)

Introduction

A patent has not only a technical function but also a legal effect. However, the legal effect of a patent is limited. This limit is in place to ensure that the patent owner can exclusively enjoy the patent's benefits in a limited area within a certain period. World Trade Organization members have agreed that the longest period that patent owners may profit from their patent should be 20 years. Once the patent owner stops paying the maintenance fee, they lose the right to monopolize the patent's benefits. Subsequently, anyone may use the patented technology to obtain benefits.

The patent maintenance system is unique in that the patent maintenance costs increase along with the length of the patent. Therefore, owners are likely to continue paying the patent maintenance fee to maintain the validity of the patent on the basis of rational decision-making considerations (Griliches, 1998; Pakes, 1986). Owners continue paying the maintenance fee as long as the patent's expected benefit is higher than this fee. The implication is that higher-quality patents have longer validity periods (Bessen, 2008; Dong et al., 2018; Maurseth, 2005; Pakes, 1986; Schankerman & Pakes, 1986).

Firm size and the stage of the technical field also affect patent maintenance length. Duguet and Iung (1997) argue that the length of patent maintenance is related to firm size and the stage of technology research and development (R&D) because companies with large-scale and strong technology R&D have advantages in human resources, financial support, and company systems. Such advantages support the realization of patented technology and the maintenance of patent validity.

Few studies have analyzed the length of patent maintenance in terms of firm size. Patent term length is often regarded as an indicator of patent quality, and it has also been verified to have a considerable analytical effect. However, the relevant application analysis aspects still require in-depth discussion. Therefore, this study examines the characteristics of firms of different sizes in terms of patent maintenance and employs mechanical engineering (ME) and electrical engineering (EE) technology as examples to compare the differences between conventional and high-tech industries.

Methodology

This study uses the patentometric method to analyze patent data. This method is widely used for R&D management, technology assessment, competitor monitoring, identification and assessment of potential sources of externally generated technological knowledge, and human resource management (Ernst, 2003).

Data collection and patent term

This study uses patent data from the United States Patent and Trademark Office (USPTO) database because the United States is one of the world's major markets. Filing an USPTO patent is a main strategic action and a meaningful symbol of global technological development for companies. It can help companies remain competitive in the global market. On the basis of the technology fields stipulated by the World Intellectual Property Organisation (WIPO), we download ME and EE technology patents granted from 1995 to 2014 from the USPTO website (USPTO, 2017b).

In the patent maintenance system of the USPTO, the longest patent term is 20 years. During the 20 years, the patent owner must pay patent maintenance fees at 3.5, 7.5, and 11.5 years after the patent was approved. If any payment is not made, the patent becomes invalid. According to these three patent maintenance fee payment times, patents can be divided into four valid durations: 4-year, 8-year, 12-year, and 20-year patents.

Firm size

Firm size in this research is obtained from the USPTO. Firms are divided into three categories according to the number of employees they have: large entities with more than 500 employees, small entities with fewer than 500 employees, and micro entities with low revenue and fewer than 500 employees. The patent fee payment standard of the USPTO is formulated to encourage innovation in small and micro entities through lower patent fees (USPTO, 2017a).

Result

Table 1 lists the number and proportion of patents obtained by firms of different sizes with varying patent term lengths for ME technology. From 1995 to 2014, large entities received 299,016 patents. Among them, 227,240 (76%) of the patents are for 20 years. In total, 28,582 (9.6%) patents are for 12 years, 30,045 (10.1%) are for 8 years, and 13,149 (4.4%) are for 4 years. Small entities received 619 patents in ME technology in 1995–2014. Among them, 390 (63.0%) of the patents are for 20 years, 67 (10.8%) are for 12 years, 98 (15.8%) are for 8 years, and 64 (10.3%) are for 4 years. Micro entities received only six patents during this period, with five (83.3%) of the patents being for 20 years and one (16.7%) being for 4 years.

Table 1. Number and proportion of patents for ME technology by differently sized firms for various patent terms

Firm size	Total patent	Patent-terms			
		4-years	8-years	12-years	20-years
large entity	299,016	13,149 (4.4%)	30,045 (10.1%)	28,582 (9.6%)	227,240 (76.0%)
small entity	619	64 (10.3%)	98 (15.8%)	67 (10.8%)	390 (63.0%)
micro entity	6	0 (0.0%)	1 (16.7%)	0 (0.0%)	5 (83.3%)
ME total	293,934	12,975 (4.4%)	29,497 (10.0%)	28,357 (9.7%)	223,105 (75.9%)

Table 2 presents the number and proportion of patents for EE technology in companies of different sizes for various patent terms. From 1995 to 2014, large entities obtained 1,187,113 patents. Among them, 1,837,003 (70.5%) are for 20 years, 107,673 (9.1%) are for 12 years, 154,641 (13.0%) are for 8 years, and 87,796 (7.4%) are 4 years. Small entities received 7,596 patents. Among them, 5,542 (72.96%) are for 20 years, 682 (8.98%) are for 12 years, 1,081 (14.2%) are for 8 years, and 291 (3.8%) are for 4 years. Micro entities received only 12 patents. Among them, 11 (91.67%) are for 20 years and the remaining 1 (8.33%) is for 8 years.

Table 2. Number and proportion of patents for EE technology by differently sized firms for various patent terms

Firm size	Total patent	Patent-terms							
		4-years		8-years		12-years		20-years	
large entity	1,187,113	87,796	(7.4%)	154,641	(13.0%)	107,673	(9.1%)	837,003	(70.5%)
small entity	7,596	291	(3.8%)	1,081	(14.2%)	682	(9.0%)	5,542	(71.0%)
micro entity	12	0	(0.0%)	1	(8.3%)	0	(0.0%)	11	(91.7%)
EE total	1,175,844	87,140	(7.4%)	152,002	(12.9%)	106,628	(9.1%)	830,074	(70.6%)

Comparison of Tables 1 and 2 reveals that the proportions of 4-year and 8-year patents of large entities for ME technology are less than those for EE technology. However, the proportions of 12-year and 20-year patents of large entities for ME technology are more than those for EE technology. Thus, large entities are more likely to maintain long-term patents for ME technology than for EE technology. In terms of small entities, the proportions of 4-, 8-, and 12-year patents are higher for ME technology than EE technology. However, the proportion of 20-year patents is lower in ME technology than in EE technology. That is, small entities are more likely to maintain long-term patents for EE technology than for ME technology. As for micro entities, the number of patents for both ME and EE technology is small. However, such entities tend to maintain long-term patents in both fields, and the proportion of EE technology patents maintained over the long term is slightly higher than that for ME technology.

Conclusions

The current results reveal that regardless of firm size, patents tend to be maintained over the long term. However, in large entities, the proportion long-term patents in ME technology is higher than that for EE technology. Small and micro entities tend to maintain more long-term patents in EE technology than in ME technology.

Acknowledgments

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 110L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by the Ministry of Science and Technology (MOST), Taiwan, under Grant No. MOST 110-2634-F-002-045- and MOST 108-2221-E-030-006-MY2.

References

Bessen, J. (2008). The value of U.S. patents by owner and patent characteristics. *Research Policy*, 37(5), 932-945.

Dong, H.-R., Chen, D. Z.,& Huang, M. H. (2018). Do long-term patents have a higher citation impact? *IEEE-IEEM 2018: IEEE International Conference on Industrial Engineering and Engineering Management*, IEEM18-P-0573.

Duguet, E. & Iung, N. (1997). *R&D Investment, Patent Life and Patent Value* (Working Paper No. G9705). Institut National de la Statistique et des Études Économiques.

Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information*, 25(3), 233-242.

Griliches, Z. (1998). Patent statistics as economic indicators: A survey. In *R&D and productivity: The econometric evidence* (pp. 287-343). University of Chicago Press.

Maurseth, P. B. (2005). Lovely but dangerous: The impact of patent citations on patent renewal. *Economics of Innovation and New Technology*, 14(5), 351-374.

Pakes, A. (1986). *Patents as Options: Some Estimates of the Value of Holding European Patent Stocks* (Working Paper No. 1340). National Bureau of Economic Research. http://www.nber.org/papers/w1340

Schankerman, M. & Pakes, A. (1986). Estimates of the Value of Patent Rights in European Countries during the Post-1950 Period. *Economic Journal*, 96(384), 1052-1076.

USPTO. (2017a, December 13). *USPTO Fee Schedule* [Text]. /learning-and-resources/fees-and-payment/uspto-fee-schedule

USPTO. (2017b, December 26). *WIPO technology fields for all patents*. PatentsView Data Download. http://www.patentsview.org/download/

Metrics Literacies:
On the State of the Art of Multimedia Scholarly Metrics Education

Isabelle Dorsch[1], Alyssa Jeffrey[2, 3], Sanam Ebrahimzadeh[4], Lauren A. Maggio[5] and Stefanie Haustein[2, 3]

[1] *isabelle.dorsch@hhu.de*
Heinrich Heine University Düsseldorf, Germany, Universitätsstraße 1, D-40225 Düsseldorf (Germany)

{ajeff069, stefanie.haustein}@uottawa.ca
[2] School of Information Studies, University of Ottawa, Canada, 55 Laurier Avenue East, Ottawa, ON K1N 6N5 (Canada)
[3] Scholarly Communications Lab, 55 Laurier Avenue East, Ottawa, ON K1N 6N5 (Canada)

[4] *sebrahimzadeh@ohri.ca*
The Ottawa Hospital, General Campus, Centre for Practice Changing Research, 501 Smyth Road, PO BOX 201B, Ottawa, ON, K1H 8L6 (Canada)

[5] *lauren.maggio@usuhs.edu*
Uniformed Services University of the Health Sciences, Bethesda, Maryland, 4301 Jones Bridge Road, Bethesda, MD, 20814-4799 (United States)

Introduction

The quantification and oversimplification of research impact and academic success is harming scholarly communities in all disciplines. Scholarly metrics, such as the h-index, impact factor, and indicators used in university rankings, are widely applied in academic tenure and funding decisions, but often inappropriately. This has created publication pressure, leading to a range of adverse effects and scientific misconduct. To improve the understanding and appropriate use of scholarly metrics in academia, Haustein (2018) proposes the concept of metrics literacies. Metrics literacies are defined as an integrated set of competencies, dispositions, and knowledge that empower individuals to recognize, interpret, critically assess, and effectively and ethically use scholarly metrics in academia.

There have been attempts to address the lack of metrics literacies in the wider academic community (e.g., *The San Francisco Declaration on Research Assessment*, the *Leiden Manifesto*, *the Metric Tide,* and *the Metrics Toolkit*). Other efforts include books and guides. Bibliometric experts have recently begun discussing metrics literacies-related concepts such as "metric wiseness" (Rousseau & Rousseau, 2015), "metric culture" (Hammarfelt & Haddow, 2018), and the "gaming" of bibliometric indicators (Derrick & Gillespie, 2013). Likewise, studies have started focusing attention on individuals' knowledge, usage, and opinions on scholarly metrics [e.g., (Derrick & Gillespie, 2013; Hammarfelt & Haddow, 2018]. Together, these initiatives represent a valuable, yet uncoordinated, first step toward educating the broader academic community on the appropriate use of scholarly metrics while the research studies provide first empirical insights. However, none of these attempts have evaluated the efficacy of those educational initiatives and, most importantly, most of them are published as textual documents, which require hours of reading. We believe that the bibliometric community needs to take responsibility of how scholarly metrics are used. The Metrics Literacies project aims to fill this gap by developing, testing, and disseminating multimedia open educational resources (OERs) to educate researchers and research administrators about scholarly metrics. A wealth of literature demonstrates the educational value of digital multimedia materials. Therefore, multimedia OERs can provide a more effective, efficient and engaging way to increase metrics literacies than using text-based resources [e.g., (Mayer, 2009)].

The Metrics Literacies project

The first part of the project focuses on the h-index, as a popular and widely-used scholarly metric, but also known for its deficiencies (Waltman & van Eck, 2012). The main objectives are to:

1. Develop multimedia OERs (e.g., videos, podcasts, infographics) to inform academics' and research administrators' understanding and appropriate application of research metrics;
2. Experimentally test and evaluate the efficacy of the educational resources and identify the most effective types of educational resources for metrics literacies topics;
3. Openly disseminate educational resources online and evaluate their uptake.

Methodology

This initiative includes three consecutive phases, which utilize a convergent mixed-methods design.

Phase 1 will include the development of the multimedia OERs by specialists. The creation of these resources, especially in regards to format selection, will be informed by the conduct of a systematic review and development of a taxonomy. These resources will include the following materials: *(1.) Five personas*[1] (four researchers and a research manager) to incorporate the use of storytelling elements (the practice of conveying information in a narrative form) and human embodiment. Specifically, incorporating a human presence, a so-called focaliser, can create authenticity and connection between the educator and the learner (Jüngst, 2010). Ensuring ethnic, cultural and gender diversity, personas come from different disciplines, different career stages, and display various levels of experience and challenges with the h-index. To make them reliable and relatable we obtain realistic examples of researchers and their publications as well as incorporate feedback from researchers which can relate in any professional way to one of the personas. *(2.) An h-index briefing note* including information on the metric (e.g., formula, computation) and evidence on the h-index issues, as for example its inconsistencies (Waltman & van Eck, 2012). The note is a textual document and based on a literature review. It builds the theoretical base for all multimedia OERs and ensures that the content of all resources remains constant. *Phase 2* will evaluate the impact of the created resources using a multi-arm randomized controlled trial (RCT), an attitude survey, a survey-based knowledge test, and semi-structured interviews. The multi-arm RCT will be composed of one control (text reading of the h-index fact sheet) and three treatment conditions (exposure of the h-index multimedia OERs). Before and after their exposure to control or treatment resources, the representative participants sample will complete an assessment. The assessment consists of the attitude survey (administered both before and after exposure) and the knowledge test on the h-index (administered as a post-test only). To establish the content validity of the knowledge test, the h-index briefing note will be used as a blueprint for the test. The semi-structured interviews will serve as means to further examine the participants' experiences with the educational resources and to understand their knowledge of the h-index. *Phase 3* will upload all educational resources online on Zenodo and YouTube and distribute them on social media to the academic community to test the engagement, popularity, and sentiment of the resources.

First outputs and next steps

Initial outputs are the first versions of the five researcher personas and the h-index briefing note (*phase 1*). The next steps encompass the completion of *phase 1*, and the implementation of *phases 2* and *3*. With our interdisciplinary project, we would like to make empirical and practical contributions to bibliometrics and multimedia education and promote metrics literacies in the academic community at large. In dependency of the evidence about most efficient OER formats and in a long-term view, the project targets the development of further educational material on the impact factor, other bibliometric indicators, altmetrics, and about the useful procession of metrics, such as field-normalized citation rates or percentile indicators. The project can be followed on Zenodo: https://zenodo.org/communities/metricsliteracies/

References

Derrick, G. E. & Gillespie, J. (2013). "A number you just can't get away from": Characteristics of adoption and the social construction of metric use by researchers. In *Proceedings of the 18th International Conference on Science and Technology Indicators* (pp. 104-116). Berlin: Institute for Research Information and Quality Assurance.

Hammarfelt, B. & Haddow, G. (2018). Conflicting measures and values: How humanities scholars in Australia and Sweden use and react to bibliometric indicators. *Journal of the Association for Information Science and Technology,* 69(7), 924-935. https://doi.org/10.1002/asi.24043

Haustein, S. (2018). Metrics Literacy: Educating Researchers and Research Support Staff Regarding Scholarly Metrics. *FORCE2018, Montreal (Canada)*.

Jüngst, H. E. (2010). *Information Comics: Knowledge Transfer in a Popular Format.* Frankfurt am Main, New York: Peter Lang.

Mayer, R. (2009). *Multimedia Learning* (2nd ed). Cambridge: Cambridge University Press.

Rousseau, S. & Rousseau, R. (2015). Metric-wiseness. *Journal of the Association for Information Science and Technology,* 66(11), 2389. https://doi.org/10.1002/asi.23558

Waltman, L. & van Eck, N. J. (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology,* 63(2), 406-415. https://doi.org/10.1002/asi.21678

[1] First version including a description and draft sketches: https://doi.org/10.5281/zenodo.4046019

How do peer reviewers (mis)interpret funding agency criteria? Results from a survey of Science Foundation Ireland grant reviewers

Thomas Feliciani[1], Junwen Luo[2] and Kalpana Shankar[3]

{[1] thomas.feliciani , [2] junwen.luo, [3] kalpana.shankar}@ucd.ie
University College Dublin, Ireland

Introduction

Although funding bodies rely extensively on peer review to evaluate grant proposals with the highest potential for success, many aspects of review processes have been criticised for lack of reliability, potential for bias, lack of transparency, and heavy reliance on overworked researchers. The literature suggests that peer review bias and unreliability can be resulted from reviewers' heterogeneity in interpreting evaluation criteria established by funding agencies (Abdoul et al., 2012; Cicchetti 1991; Hug and Aeschbach 2020; Lee et al., 2013; Pier et al., 2017). To examine this issue more closely and as part of a larger project on grant peer review, we designed and administered an online survey of highly experienced reviewers for Science Foundation Ireland (SFI), Ireland's largest funding agency, to learn more about their experiences and perceptions of grant peer review criteria. In this poster, we describe some survey results, focusing on reviewers' perceptions of the evaluation criteria on a typical review form.

Survey design

We began with a content analysis of 527 peer review forms of two SFI funding programmes (2014-2017) and identified the twelve topics most often commented on by the reviewers (listed in Figure 2). The twelve topics were then included in two questions on our survey of the same reviewer pool:

- Which topics/criteria do you consider when evaluating the three review sections (applicants, proposed research, and potential for impact)? [Tick all topics that apply]
- In general (i.e. not just for SFI), how do these topics/criteria weigh in your evaluation of the "Impact" section of a proposal? [5-point scale from "not at all important" to "extremely important"]

The survey methodology and questionnaire are reported in full in Shankar et al. (2020). To ascertain the level of experience of SFI reviewers, respondents were asked when they had received their highest academic degree and approximately how many reviews they performed for any funding body. Other survey themes include questions about the procedures by which proposals are evaluated (clarity of review guidance, clarity of evaluation scale, time involved to review), the process of conducting the evaluation (interpretation of the agency's evaluation criteria, interpretation of the grading scale, group dynamics), the inclusion of non-academic reviewers, and transparency of reviewers' identities. There were 36 questions following an informed consent sheet. Question types were phrased as either 5-point Likert scales or open-ended, some questions with spaces for comment.

Survey respondents

The survey link was sent in June 2020 to all 1591 reviewers involved in our addressed funding programmes. The survey was kept open through the end of July 2020. 310 respondents completed the survey. Figure 1 shows the demographic composition of the survey respondents compared with the general demographics of SFI reviewers from 2014 to 2017 (the time period of our study). The funding agency relies exclusively on an international reviewer pool including researchers from academia and industry. Because reviewers are anonymous, SFI sent out the survey on our behalf. SFI did not provide any substantive comments or feedback on the survey, nor were they privy to the identities of individuals who took the survey.

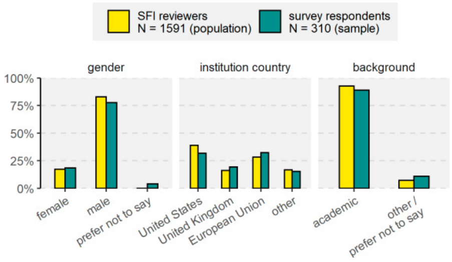


Figure 1. Survey demographics.

Evaluation criteria: Not so clear after all

Like many other funding agencies, SFI provides reviewers with review forms structured in separate sections, each section being dedicated to the evaluation of a specific evaluation criterion, i.e. "potential for impact". Funders also provide instructions as to what should be considered when evaluating these criteria. In the case of SFI, the evaluation criteria and corresponding instructions were roughly consistent through the time period and across the funding programmes we examined.

Despite these instructions, scholars have noted how there is often variation between reviewers in what aspects, or topics, reviewers focus on when evaluating a given criterion (see e.g. Langfeldt, 2001). In order to understand the consequences of such interpersonal variation, we first set out to find whether and to what degree there is variation in the interpretation of the evaluation criteria.

Figure 1 shows the answers to these two questions for the criterion "potential for impact": the plot shows the *frequency* with which each topic was considered for the evaluation of that criterion, and its deemed *importance*. Results show that some of the topics are reported more often than others (with higher frequency). Curiously, the topics that reviewers consider more often (in yellow) are not necessarily the ones deemed to be more important (higher value on the X axis): for example, the topic "research design" is considered of relatively high importance for the evaluation of "potential for impact"; however, it is among the least reported topics. This observation alone shows some degree of dissonance between which topics reviewers claim to consider when evaluating a specific criterion and which they deem the most important.

Furthermore, for all the topics we found large variance between reviewers in both questions. This is illustrated, e.g., by the length of the boxplots in Figure 2: for some topics like "mitigating risk", many responses ranged all the way from "not at all important" to "extremely important". This signals poor inter-reviewer agreement on which topics are considered and which ones are important for the evaluation of a proposal's potential for impact.

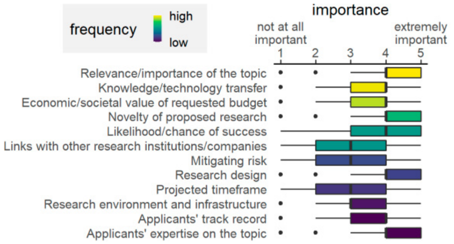


Figure 2. Review topics: frequency and importance for the evaluation criterion "potential for impact".

Future Research

29 survey respondents (9%) left email addresses to arrange follow-up interviews. We conducted 16 interviews and are in the process of analysing the data to explore further the reviewers' perceptions on evaluation criteria and other topics (e.g. grading scales). We will use this data to support simulation studies of reviewers' variant interpretations of evaluation criteria and how such variances would influence inter-rater reliability in peer review.

Conclusion

In this poster we empirically show that reviewers' diversity in the interpretation of evaluation criteria may constitute one reason for poor inter-rater reliability. Our results suggest that to some degree reviewers do what they want in spite of provided instructions. While more research in other funding programmes and agencies is clearly needed, we would point to a need for attention to this reason for any policy changes in grant peer review.

Acknowledgments

This material is based upon works supported by the Science Foundation Ireland under Grant No.17/SPR/5319. We also thank Dr Pablo Lucas and Dr Lai Ma for their inputs.

References

Abdoul, H., Perrey, C., Amiel, P., Tubach, F., Gottot, S., Durand-Zaleski, I. & Alberti, C. (2012). Peer review of grant applications: Criteria used and qualitative study of reviewer practices. *PLOS ONE* , 7, e46054.

Cicchetti, D.V. (1991). The reliability of peer review   for manuscript and grant submissions: a cross-disciplinary investigation. *Behavioral and Brain Sciences,* 14, 119-135.

Hug, S. E. & Aeschbach, M. (2020). Criteria for assessing grant applications: a systematic review. *Palgrave Commun,* 6, 37.

Langfeldt, L. (2001). The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome. *Social Studies of Science*, 31(6), 820-841.

Lee, C. J., Sugimoto, C. R., Zhang, G. & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1), 2-17.

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., Ford, C. E. & Carnes, M. (2018). Low agreement among reviewers evaluating the same NIH grant applications. *PNAS*, 115(12), 2952-2957.

Shankar, K., Luo, J., Ma, L., Lucas, P. & Feliciani, T. (2021). *SPRING 2020 survey: peer review of grant proposals* (Dataset). FigShare. https://doi.org/10.6084/m9.figshare.13651058.v1

Impact of Secretariat Research Groups in Brazil

Rosimeri Ferraz Sabino[1], Fabio Gomes Rocha[2] and Alejandro C. Frery[3]

[1] *rf.sabino@gmail.com*
Federal University of Sergipe, São Cristóvão (Brazil)

[2] *gomesrocha@gmail.com*
Universidade Tiradentes/Lisorpe and ITP, Aracaju (Brazil)

[3] *alejandro.frery@vuw.ac.nz*
Victoria University of Wellington, School of Mathematics and Statistics, Wellington (New Zealand)

Introduction

The sharing of Brazilian scientific knowledge counts on the Directory of Research Groups (DGP[1]) support, a project created in 1992 by the National Council of Scientific Development (CNPq). The DGP's objectives are to promote the exchange of information between the scientific and technological communities, to provide information for the management of science and technology activities, and to preserve the memory of Brazilian scientific-technological activities (CNPq, 2020). Researchers identify, analyze, and discuss and disseminate knowledge, promoting scientific production (Aman, 2019). Such production is classified before Qualis[2], an evaluation system for the Brazilian postgraduate system through scientific journals' stratification (Rocha, Sabino, and Frery, 2020). In this sense, this work aims to characterize the scientific production of Secretariat research groups in Brazil, observing productivity and impact.

Materials and Methods

We collected data carried out in December 2020, from the current DPG base, with the parameters: "secretariat" in the fields "Name of the group," "Name of the research line," and "Keyword of the research line," and "Certified" and "Not updated" in the field "Situation." This last field identifies if the group has been updated in the previous twelve months. This search resulted in 15 groups, with each research line analysed to identify the relationship's consistency with the Secretariat field. With this, we identified a group that did not present such a relationship, and we excluded it from the search. The 14 groups investigated added up 243 members and 74 lines of research, of which only 29.21% are related to Secretariat. Table 1 shows the groups, the date of their creation, and number of members.

Table 1. Secretariat research groups in Brazil

ID	Group	Data	Members
1	Grupo de Pesquisa em Secretariado Executivo Bilíngue	2002	15
2	Gestão do Conhecimento nas Ciências Sociais Aplicadas	2009	30
3	Grupo de Pesquisas Interdisciplinares em Secretariado	2009	25
4	Núcleo de Pesquisa de Estudos em Secretariado Executivo e áreas afins	2011	16
5	Grupo de Estudos e Pesquisas em Secretariado Executivo	2014	18
6	Núcleo Interdisciplinar de Estudos em Secretariado Executivo	2014	5
7	Núcleo de Pesquisas Aplicadas em Gestão, Secretariado Executivo e Economia	2014	31
8	Gestão, Assessoria Executiva, Secretariado e Sociedade	2016	23
9	Grupo de Pesquisas em Práticas Secretariais	2016	14
10	Observatório Latino-Americano de Pesquisa em Secretariado Executivo	2016	16
11	Pesquisa e Prática em Gestão e Secretariado	2016	8
12	Grupo de Estudos em Secretariado Executivo	2017	13
13	Estudo Multidisciplinar de Gestão	2018	9
14	Grupo Interdisciplinar Latinoamericano de Estudos e Pesquisa em Secretariado Executivo	2018	20

[1] http://lattes.cnpq.br/web/dgp

[2] https://sucupira.capes.gov.br/sucupira/

We kept the original groups' names in Portuguese so the reader may find them in the DGP data base.

Based on the groups mapping, we defined the following research questions: Q1 -How is the scientific production of Secretariat research groups in Brazil distributed quantitatively? Q2 - What is the distribution of this product in the Qualis strata? Q3 - What is the impact of the research groups investigated? We verified each researcher's production in each group to answer these questions. As selection criteria, we define the terms "secretary," "secretaries," and "secretariat" in the title or abstract of articles in the Lattes Platform[3]. Then, we identified the Qualis stratum of the journals using the ChromeQualis tool. Finally, we used the "Publish or Perish" tool, with Google Scholar, to verify the number of citations of the publications to measure the group's impact (H-index).

Results

In response to Q1, we obtained 184 articles published in the period from 2013 to 2020. The groups are identified in the figures below by the corresponding ID in Table 1.

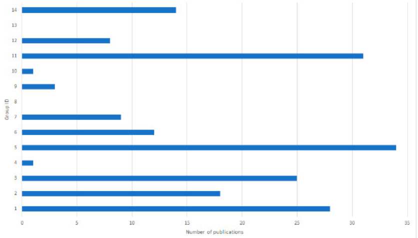


Figure 1. Number of publications per research group, from 2013 to 2020

We analysed the publications in two blocks: Qualis 2013-2016 ("A1", "A2", "B1" to "B5" and "C") and Qualis 2017-2020 ("A1" to "A4", "B1" to "B4" and "C"), the latter being in a preliminary version. In both blocks, the publications are concentrated in strata "B1," "B2," and "B3", the lowest in the rank, as shown in Figures 2 and 3.

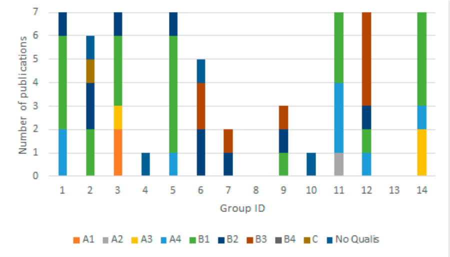


Figure 2. Distribution of publications in Qualis 2013-2016 strata

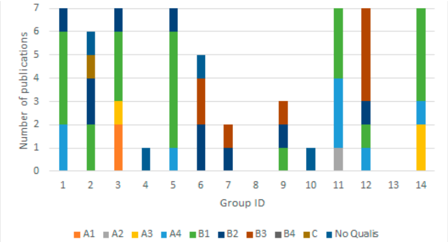


Figure 3. Distribution of publications in Qualis 2017-2020 strata

We analysed the impact of the papers produced in the groups. The largest H-Indexes (6 and 4) were from groups 1, 2, and 3. These are the oldest groups in operation. Of the recently created groups, only two (#13 and #14) have H-Index larger than zero, as shown in Figure 4.

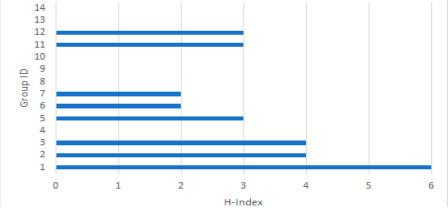


Figure 4. H-Index

Discussion and conclusion

The scenario of research groups in Secretariat indicates significant expansion since the decade of 2010, but with little progress compared to publications in higher strata. From 2013 to 2016, there are only two publications in the "A2" stratum, concentrated in a single group. From 2017 to 2020, there are only three publications in the "A1" and "A2" strata, distributed in two groups. On the impact of the productions, we found that only eight groups have H-Index. Although the groups' time of existence has influence in the productions' impact, some groups do not have any publication since their creation. We conclude that the output of Secretariat is concentrated in a few research groups, with negative impact in the evolution of the field and the National contribution to the scientific community.

References

Aman, V. (2019). Internationally mobile scientists as knowledge transmitters – a lexical-based approach to detect knowledge transfer. *Proceedings of the 17th ISSI,* (pp. 2199-2208). Roma: Sapienza University of Roma.

CNPq. *Objectives.* Retrieved December 2, 2020, http://lattes.cnpq.br/web/dgp/objetivos/

Rocha, F. G., Sabino, R. F. & Frery, A. C. (2020). Analysis of the international impact of the Brazilian base "Qualis"-Education. *Scientometrics*, 135(3), 1949-1963.

[3] http://lattes.cnpq.br

Altmetrics for evaluation of medical research in Germany

Nicholas Fraser[1], Paula Bräuer[1,2] and Isabella Peters[1,2]

{n.fraser, p.braeuer, i.peters}@zbw.eu
[1] ZBW – Leibniz Information Centre for Economics, Kiel (Germany)
[2] Kiel University, Kiel (Germany)

Introduction

Since 2004, all medical faculties in Germany have been partially allocated funding according to performance indicators based predominantly on two scientometric criteria: (1) the amount of awarded third party funding, and (2) the number and quality of authored publications. Whilst the exact model by which each medical faculty evaluates their own publication performance varies, the evaluation of publication 'quality' has largely been based on citation-based metrics, namely Journal Impact Factors (JIF). Such JIF-based measures have been widely as indicators of individual publication quality, both nationally by the Association of Scientific Medical Societies in Germany (Hermann-Lingen et al., 2014), and internationally through initiatives such as the San Francisco Declaration on Research Assessment (DORA; https://sfdora.org/).

An additional factor complicating the usage of citation-based metrics for medical research assessment relates to an apparent citation "preference" or "advantage" of basic research (i.e. studies of fundamental functions and systems) in comparison to clinical research (i.e. studies of health and disease treatment in human subjects; van Eck et al., 2013; Donner & Schmoch, 2020; Ke, 2020). Applying citation-based metrics at an institutional level must therefore take into account differences in the research focus of each individual institution.

Altmetrics are metrics that capture countable signals for the access, usage and sharing of research objects on online platforms. They can provide a measure of public interest or discussion of scholarly works (Tahamtan & Bornmann, 2020), which may be an important contribution to multidimensional research evaluation methods, particularly in biomedical and health science fields which have been found to rank highly in terms of sharing rates on social media and news platforms (Costas et al., 2015).

In this poster, we will present results of an investigation into how articles authored by researchers at German medical research institutions are shared on various online platforms. In doing so, we will also assess how individual altmetric indicators vary with respect to their tendencies towards research levels (basic vs clinical research).

Methods

A list of titles and ISSNs for journals indexed in MEDLINE (N = 5007), a biomedical bibliographic database maintained by the US National Library of Medicine, were downloaded. Journals were matched to those indexed in the Web of Science (WoS), leveraging the data infrastructure of the German Competence Centre for Bibliometrics (http://www.forschungsinfo.de/Bibliometrie/en/index.php), on the basis of exactly-matching titles or ISSNs. We excluded journals with the WoS classification of "Multidisciplinary Sciences"; 16 journals), which included journals with a non-exclusive biomedical focus such as *PLOS ONE* or *Scientific Reports*. In total 4,442 MEDLINE journals were matched to journals in WoS.

We subsequently extracted publication metadata (DOI, publication year, article title, abstract) for all articles published in these journals with at least one author associated with a German research institution. Articles were limited to those published between 2012 and 2018, to "Article" and "Review" types, and to those with a valid DOI. In total we extracted details of 336,193 articles.

Altmetrics information were extracted from Altmetric (https://altmetric.com), by iteratively querying the API for each article DOI. We extracted counts from 5 main sources: Twitter, Facebook, mainstream media, blogs and policy documents (which include documents issued from government guidelines, reports or white papers; independent policy institute publications; advisory committees on specific topics; and international development organisations[1]). The Altmetric API only provides a valid response when an article has been mentioned in at least one of the single sources tracked – thus queries resulting in invalid responses ("Not Found") were included with counts of 0 for all sources considered.

To understand how altmetrics vary by indicators and research levels, we rely on visualisations of term co-occurrence maps using the *VOSViewer* software (van Eck & Waltman, 2010). In brief terms, each node in the map represents a term, whereby the size of the node is proportional to the total number of times a term is mentioned in the title and abstracts of our set of articles, and the distance between the nodes is proportional to the number of times that terms co-occur together in the same document.

Preliminary Results

Figure 1 shows three term maps as an indicator of our preliminary results – full results will be presented in the conference poster. Panel A shows clustering of terms present in our sample of articles – notably we see a transition from terms that we consider to align with basic research (e.g. “cell”, “protein”, “property”, “structure”) on the left side (red), to terms that we consider to align with clinical research (e.g. “patient”, “therapy”, “diagnosis”, “participant”) on the right side (blue). Panel B replicates Panel A in structure, but differs in that colors represent the strength of mentions of a term on Twitter (darker red = more-tweeted terms). We observe a slight tendency of articles containing clinical-related terms to be more tweeted than articles containing basic-related terms. In Panel C, colors refer to the number of citations in policy documents. We observe a stronger tendency for articles containing clinical-related terms to be cited in policy documents. The results highlight variation in the response of individual metrics to different research levels in medical research; for conducting evaluation of research at the institutional level, understanding these differences will be of key importance.

Future work will expand on these preliminary results, by considering further factors influencing these relationships, such as author and publication properties, or collaboration networks.

Acknowledgements

The authors acknowledge funding from the German Federal Ministry for Education and Research, grant number 01PU17011D (QuaMedFo).

References

Costas, R., Zahedi, Z. & Wouters, P. (2015). Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10), 2003-2019.

Donner, P. & Schmoch, U. (2020). The implicit preference of bibliometrics for basic research. *Scientometrics*, 124(2), 1411-1419.

Herrmann-Lingen, C., Brunner, E., Hildenbrand, S., Loew, T. H., Raupach, T., Spies, C., Treede, R.-D., Vahl, C.-F. & Wenz, H.-J. (2014). Evaluation of medical research performance – position paper of the Association of the Scientific Medical Societies in Germany. *GMS German Medical Science,* 12(11).

Ke, Q. (2020). The citation disadvantage of clinical research. *Journal of Informetrics*, 14(1), 100998.

Tahamtan, I. & Bornmann, L. (2020). Altmetrics and societal impact measurements: Match or mismatch? A literature review. *El Profesional de La Información*, 29(1), e290102.

van Eck, N. J. & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538.

van Eck, N. J., Waltman, L., van Raan, A. F. J., Klautz, R. J. M. & Peul, W. C. (2013). Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research. *PLoS ONE*, 8(4), e62395.

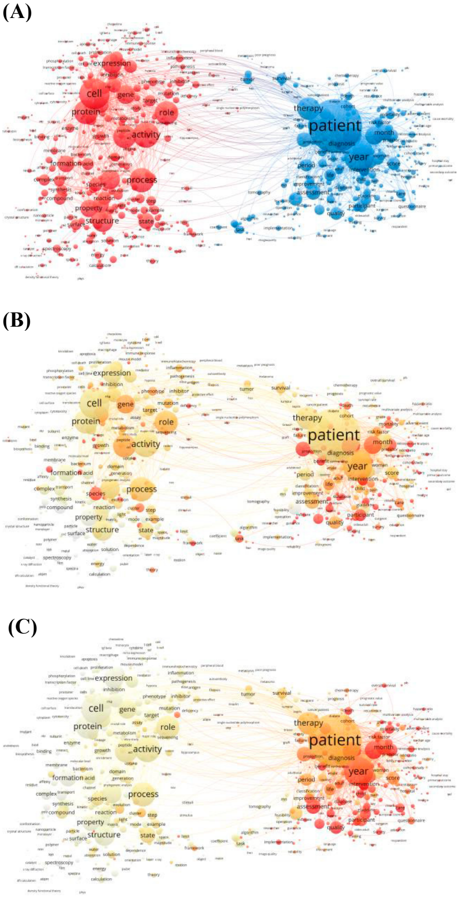


Figure 1. Term co-occurrence maps generated with VOSViewer (term frequency: > 250; term relevance: 60%) (A) Map overlain by topical clusters. (B) Map overlain by Twitter strength (darker red = terms more tweeted). (C) Map overlain by policy-document strength (darker red = terms cited more in policy documents).

[1] https://help.altmetric.com/support/solutions/articles/6000236695-policy-documents

The three-dimensional activity index

Joel Emanuel Fuchs

jfuchs@uni-wuppertal.de
University of Wuppertal, Institute of Sociology, Gaußstraße 20, D-42119 Wuppertal (Germany)

Introduction

In sociological data analysis the comparison of data from different countries or institutions occurs all the time. The comparison of absolute values is often problematic due to the different sizes of the observed entities. One solution for comparing different sized entities is the activity index (AI). It enables the normalized international or inter-institutional contrasting of various fields. Although the AI is a long-used instrument, it lacks self-specific instruments to analyse itself. In this paper, we first want to present the AI. After that, we will introduce a new measure called the three-dimensional activity index (3D-AI) motivated by the statistical expected value. In the last part we will show how to use the 3D-AI to centre the basic activity index.

Used Data

For the sake of comprehension, data is used to visualise the new indices. Data from the European Patent Office (EPO) is open, easy to understand and traceable. So, we decided to use the granted patents per field of technology and per country of residence for 2011-2015 available from https://www.epo.org. The data is smoothed by a 3-year binomial filter to visualize the field and country specific trends better.

From the dataset follows, that the basic population consists of 46 countries plus 1 residual category divided into 35 fields over five years. All 8,225 data points will be used for calculation, but to keep the visual analysis clear, only the two countries Germany and the United Kingdom (UK) and the two fields 'Food chemistry' (FC) and 'Semiconductors' (SC) will be represented. This choice is arbitrary; the focus lies on the formulae presented later.

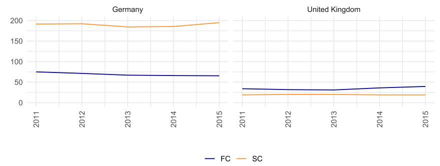


Figure 1. Granted patents.

The activity index

Of course, we can see in figure 1, that the shares of FC and SC of the overall granted patents must be more similar to each other in the UK than in Germany. But 'seeing' or comparing the absolute values is too intangible. Therefore, a relative index is often used, which directly reveals such differences between countries regarding the underlining fields. We will call it the activity index (AI), as denominated by Narin et al. (1987). But it is also known under the revealed technological advantage (Soette & Wyatt, 1983), revealed comparative advantage (Balassa, 1965) or the Balassa index (Ibid.).

Let x_{ijt} be the granted patents of country j regarding the field i in the year t. The AI relates the share of one field of a country $\left(x_{ijt}/\sum_i x_{ijt}\right)$ to the share of the same field but of all countries $\left(\sum_j x_{ijt}/\sum_{ij} x_{ijt}\right)$. So, we get

Formula 1. Activity index.

$$AI_{ijt} := AI\left(x_{ijt}\right) := \frac{x_{ijt}/\sum_i x_{ijt}}{\sum_j x_{ijt}/\sum_{ij} x_{ijt}}.$$

We calculated the AI for all 8,225 data points, but in figure 2 we will again only show Germany and the UK as well as FC and SC. Because of the different sums used for the AI, it is important to mention which values were calculated and which data points were used.

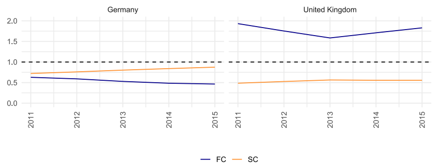


Figure 2. Activity index.

Figure 2 shows the AI corresponding to figure 1. The dashed line represents the average across all combinations of country and field. It is obvious that Germany is closer to the average than the UK. The second observation is that, except of Semiconductors in Germany, all other fields veer away from the average over time.

The temporal activity index

The values of the AI are calculated year by year. This is done, because all 1,645 data points of one year are integrated into the calculation of a single AI value by the composed sums. If we want AI values, that do not depend on the year, we could summarise all values by country and field over all five years. We get an AI that is constant over time so we will denote it the Temporal Activity Index (TAI).

Formula 2: Temporal activity index.

$$TAI_{ijt} \coloneqq TAI(x_{ijt}) \coloneqq \frac{\sum_t x_{ijt} \,/\sum_{it} x_{ijt}}{\sum_{jt} x_{ijt} \,/\sum_{ijt} x_{ijt}}.$$

Arithmetic mean

Let us calculate the arithmetic mean (AM), but not for all AIs, just for the AIs of a single country and a single field. Therefore, y shall be the number of years. Then we get $AM(AI_{ijt}) = \frac{1}{y}\sum_t AI(x_{ijt})$. This describes the AM of the AI values. We could also calculate the AI values of the arithmetic means of each part of the AI, so these would be $AM(x_{ijt}) = \frac{1}{y}\sum_t x_{ijt}$, $AM(\sum_i x_{ijt}) = \frac{1}{y}\sum_{it} x_{ijt}$, $AM(\sum_j x_{ijt}) = \frac{1}{y}\sum_{jt} x_{ijt}$ $AM(\sum_{ij} x_{ijt}) = \frac{1}{y}\sum_{ijt} x_{ijt}$. Using these AMs for calculating the AI, we will get the TAI as aforementioned. We conclude that the TAI is very similar to the AM.

Table 1. Temporal activity index.

Country	FC	SC
Germany	0.535	0.796
United Kingdom	1.759	0.535

The three-dimensional activity index

What does this have to do with the 3D-AI? The 3D-AI is the fraction of the classic AI and the TAI, as

Formula 3. Three-dimensional activity index.

$$3D\text{-}AI(x_{ijt}) \coloneqq \frac{AI(x_{ijt})}{TAI(x_{ijt})} = \frac{x_{ijt} \cdot \sum_{ij} x_{ijt} \cdot \sum_{it} x_{ijt} \cdot \sum_{jt} x_{ijt}}{\sum_{ijt} x_{ijt} \cdot \sum_t x_{ijt} \cdot \sum_j x_{ijt} \cdot \sum_i x_{ijt}}.$$

So, the 3D-AI centres the AI by its average over the years represented by the TAI. We therefore improve the AI by disadvantaging other parts of it, as we can see in figure 3.

The first two indices of the AI can be interchanged, i.e. $ai(x_{ijt}) = ai(x_{jit})$. We denote the newest value the 3D-AI, because all three indices can be interchanged, i.e. $ai(x_{ijt}) = ai(x_{tij}) = ai(x_{itj})$ and so on.

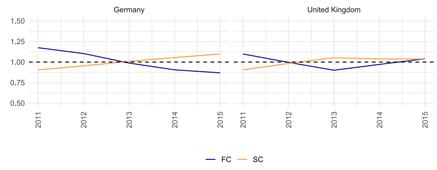


Figure 3. Three-dimensional activity index.

So, what are the advantages and disadvantages? Because we only divide the lines from figure 2 by a constant, the centred 3D-AI values are very similar, as we would expect from centred values. But now we can directly see in which year which country regarding which field performs like its average. For example, UK's FC line intersects with the dashed line about the year 2012 and 2014/2015. Before 2012, UK's FC performed above its own average, same after 2014. This is new information, which the classic AI and the absolute values could not express. The 3D-AI does not replace the classic AI, because it has some disadvantages. Due to the centralisation we cannot measure which field performs above average and which below. There are some more disadvantages, which would go beyond the scope of this paper.

In the future, there is a lot more to do. The TAI can be a better mean for the AI, but for a complete centring we also need an (empirical) variance. Indeed, if we want to analyse data by the activity index, we should think about reconstructing all empirical instruments, not only the mean and the variance, but also differentiation, correlation and so on. Perhaps we would benefit from a whole new toolbox designed especially for inter-institutional analyses.

Acknowledgments

We thank Jennifer Madine for proofreading this paper. Further we thank the European Patent Office for their free data providing service. All graphics and calculations were made by the help of R, RStudio and the huge package service around R, especially magrittr, data.table and ggplot2.

References

Balassa, B. (1965). Trade liberalization and "revealed" comparative advantage. *The Manchester School of Economic and Social Studies,* 32, 99-123.

Narin, F., Carpenter, M. P. & Woolf, P. (1987). Technological assessments based on patents and patent citations. In Grupp, H. (Ed.), *Problems of measuring technological change*, Cologne, 107-119.

Soete, L. G. & Wyatt, S. M. E. (1983). The use of foreign patenting as an internationally comparable science and technology output indicator. *Scientometrics,* 5, 31-54.

Reasons for Retraction Affect the Impact on Citations

John Gibson[1] and Keruma Gibson[2]

[1] *jkgibson@waikato.ac.nz*
University of Waikato, Private Bag 3105, Hamilton 3240 (New Zealand)

[2] *kgg18@uclive.ac.nz*
University of Canterbury, Private Bag 4800, Christchurch 8140 (New Zealand)

Introduction

The retraction of articles from peer reviewed journals is a growing issue. Retraction notices may be issued for what can be considered as the ‘bad behaviour’ of authors, such as violating ethics and privacy requirements, (self-) plagiarism and other manipulation of citations, missing credit and other authorship issues, and interfering with the peer review process. While bad behaviour could be due to attempts by authors to shield unreliable results from critical scrutiny, there is no inherent reason to consider results of studies to be wrong, just because the authors engaged in bad behaviour.

In contrast, for retractions due to either fabrication or falsification of data, or for errors in the analysis or the data, there are grounds to disbelieve results of the study. To date, the literature on the impact of retractions on subsequent citations (e.g., Furman, Jensen & Murray, 2012) which finds post-retraction citations decline over 60% relative to the citations for a matched control sample, does not examine if the impacts depend on the reason for retraction. On the other hand, research that looks at the reasons for a retraction, such as Cox, Craig & Tourish (2018), does not estimate the impacts of the retraction.

In order to fill this gap in the literature, we examine the impacts of retraction, for articles in psychology journals. We distinguish between retractions due to (a) bad behaviour of authors, and (b) retractions due to data fabrication, falsification and errors, which is a set that we group together as ‘dodgy results’.

Research Design

The analysis is based on 402 articles in psychology journals. The ‘treatment group’ is 143 articles that were retracted; a subset of 160 articles studied by Craig et al. (2020). The reason for not using all articles is that some were retracted due to errors by the journal, and some were not available in *Web of Science* (WoS). The control group is 259 nearest neighbour articles, which were published immediately before or after the retracted article in the same issue of a journal (if a retracted article was first or last in an issue, it only has one nearest neighbor, which is why therefore the control group does not have n = 286). This research design controls for the age pattern of citations, for journal impact factors, and for the effect of position within a journal issue on citations.

The nearest neighbor articles give a counterfactual to what the retracted article’s citations might have been had it not been retracted. One threat to use of these articles as a counterfactual is that they may be affected by the retraction of their neighbour, but there is evidence to suggest that such spillovers are rare (Azoulay et al., 2015). Citations for all 402 articles (including year of each citation) were retrieved from Web of Science (WoS) and Google Scholar (GS), totaling nearly 20,000 and 50,000, respectively.

Results

Articles retracted because of the bad behaviour of authors (e.g. due to plagiarism) had accrued the same citations (by early 2021), on average, as their control group articles (Figure 1a). In contrast, articles that were retracted because they had dodgy results (due to data fabrication, errors in analysis and so on) had accrued only one-quarter as many citations as their control group articles (Figure 1b).

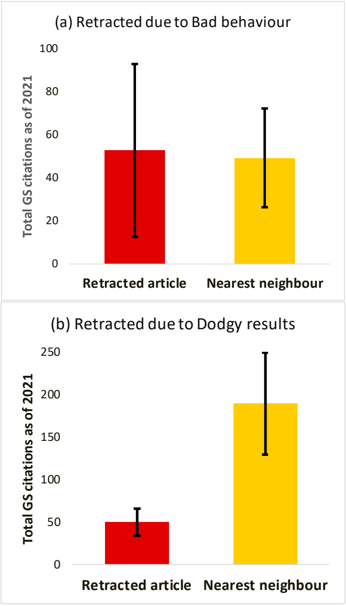


Figure 1. Mean citations (GS in early 2021) for retracted articles and their control groups

Table 1. OLS and median regression estimates of gap in citations between control group articles and retracted articles, by reason for retraction.

	Mean Gap in Citations		Median Gap in Citations	
Reason	**WoS**	**GS**	**WoS**	**GS**
Bad behaviour	-54.68	-144.80	-20.95	-47.67
	(2.69)	(2.47)	(1.79)	(1.75)
	[-0.21]	[-0.20]		
Dodgy results	49.06	135.45	21.35	49.41
	(2.65)	(2.55)	(2.01)	(2.10)
	[0.21]	[0.20]		

Notes: The gap is defined as citations for the control group articles minus those for the retracted article, WoS is Web of Science and GS is Google Scholar. $N = 143$, with t-statistics in ( ) and standard deviation effects in [ ]. Each cell is from a separate regression, with controls for article age and time since the retraction notice was issued also included.

We calculate the gap in citations (as of early 2021) as the amount by which citations for control group articles exceed those for retracted articles. The gap averages 46 (WoS) or 109 (GS), and is larger than two-thirds of mean citations for the control group. The average gap obscures a key difference; the gap varies with the reason for the retraction. We focus on retractions due to bad behaviour (n = 26) and due to dodgy results (n = 109); the other n=8 retractions give unclear or vague reasons for the retraction.

Regression results in Table 1 show that the gap in citations between retracted articles and their matched controls is significantly smaller, by about 0.2 standard deviations, if the retraction is because of bad behaviour reasons. In contrast, the gap is significantly larger if the reason given for the retraction is because of things like data fabrication or errors in the analysis, which we group together as ‘dodgy results’ reasons.

These patterns show up with both GS and WoS as the source of citations data. The patterns also show up when median regressions are used, to account for the skewed nature of citations. In all cases, the regressions in Table 1 control for the article age and time since retraction.

A key question is when does the gap in citations between retracted articles and the matched controls appear. We answer this using WoS citations, as the timing of citations is more easily obtainable from this source. However, given the similarity of results when using GS and WoS in Table 1, the patterns we find should hold more broadly.

For the analysis of these timing effects we restrict attention to the retractions that are due to ‘dodgy results’, as it is only for these that there is a gap in citations (as of early 2021) between the retracted article and the matched controls (Figure 1). Most of this gap in citations occurs after the retraction notice is published (Figure 2). For the three years before the retraction notice, the yet-to-be retracted articles averages two-thirds as many citations per year as their matched controls, while for the three years after, the retracted articles get only 14% of the citations per year of their matched controls. A gap begins to open the year ahead of the retraction notice, suggesting that the findings of the retracted article were already beginning to be discounted (the pattern holds if we weight neighbours by keyword overlap).

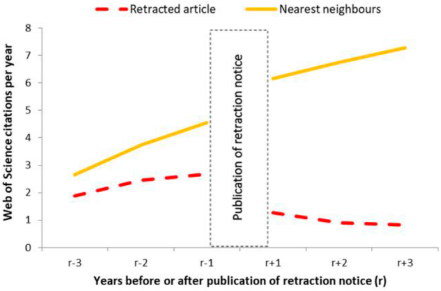


Figure 2. Average citations per year (in WoS), for three years before and after publication of the retraction notice, for subsample of articles retracted due to ‘dodgy results’ reasons.

Conclusions

Retractions of articles in psychology journals have a substantial effect on citations but only when the reason for retraction relates to unreliable findings, stemming from either fabrication or falsification of data, or from errors in analysis or data. If retraction is for other reasons, such as unethical behaviour of the authors, there appears to be no penalty through reduced citations.

Acknowledgments

We are grateful to Dr Adam Cox for sharing a list of retracted articles in psychology journals and to the University of Waikato Summer Research Scholarships program for funding support.

References

Azoulay, P., Furman, J. L., Krieger, J. L. & Murray, F. (2015). Retractions. *Review of Economics and Statistics*, 97(5), 1118-1136.

Cox, A., Craig, R. & Tourish, D. (2018). Retraction statements and research malpractice in economics. *Research Policy*, 47(5), 924-935.

Craig, R., Cox, A., Tourish, D. & Thorpe, A. (2020). Using retracted journal articles in psychology to understand research misconduct in the social sciences: What is to be done? *Research Policy*, 49(4), 103930.

Furman, J. L., Jensen, K. & Murray, F. (2012). Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. *Research Policy*, 41(2), 276-290.

Altmetrics System for the Evaluation of Chinese Scholars

Guo Ying[1] and Xiao Xiantao[2]

[1]*guoying49@whu.edu.cn*
School of Information management, Wuhan University, Wuhan 430079 (China)

[2]*xxt@lzb.ac.cn*
Lanzhou Information Center, Chinese Academy of Sciences, Lanzhou 730000 (China)

Introduction

Altmetrics was proposed by Preim (2010) in a website declaration, which defined altmetrics as "the creation and study of new metrics based on the social Web for analyzing and informing scholarship". Since then, many scholars have tried to apply altmetrics to evaluate the impact of articles, books, and journals. However, only a few of scholars have focused on author-level altmetrics. Based on the prior literature and the various tools (Altmetric.com, PlumX, et al.) of altmetrics, we summarize the platforms where the altmetrics indicators are usually drawn, and divide these platforms into five categories (See Table 1). It is found that most of the platforms are internationally popular. However, due to the policy restrictions in China and the language differences between international English and Chinese, these platforms are either inaccessible or not widely used by Chinese scholars. Hence, using indicators from them to evaluate Chinese scholars, there will be huge biases in the results of the evaluation.

Table 1. The common platforms where altmetrics indicators are originated

Categories of platforms	Examples of platforms	Examples of indicators
Online library	Scopus, PLoS One	downloads, views
Academic social platforms	Research Gate, F1000, Mendeley	Mentions, views, and likes
Comprehensive social platforms	Facebook, Twitter	Mentions, likes, comments
Encyclopedia platforms	Wikipedia	Mentions, views
News websites	CNN	Mentions

Accordingly, this paper explores the indicators in platforms which are similar with the internationally popular platforms and popular in Chinese scholars at the same time, and tries to build an altmetrics system to evaluate Chinese scholars scientifically.

Methodology and data

Based on the categories of the commonly used platforms, we investigate similar representative platforms which are popular in Chinese scholars. After investigation, we select CNKI, Sciencenet.cn, Wechat and Weibo, and Baidu baike as the representatives of online library, academic social platforms, comprehensive social platforms, and encyclopedia platforms, respectively. Besides, we choose People's Daily Online, Xinhuanet, and China.org.cn, which are the Top 3 of the 2017 news website communication power ranking of China as the representatives of news websites. On the basis of the altmetrics indicators adopted in the prior literatures, 16 new variables are drawn from the platforms above (See Table 2).

Table 2. The selected platforms and variables

Platforms	Variables
CNKI	papers (*CNKI_P*), downloads (*CNKI_Dl*)
Sciencenet.cn	mentions (*Snet_M*), views (*Snet_V*), recommendations (*Snet_R*), comments (*Snet_C*)
WeChat	mentions in WeChat official accounts (*WeCh_M*)
Weibo	mentions (*Wb_M*), reposts (*Wb_R*), comments (*Wb_C*), likes (*Wb_L*)
Baidu baike	views (*BD_V*), likes (*BD_L*), shares (*BD_S*)
News websites	mentions by all news websites (*News_All*), mentions by the Top3 representative news websites (*News_Rep*)

Given the large number of indicators, we use the principal component analysis (PCA) to reduce the number of dimensions. And the receiver operator characteristic (ROC) curve is used to test the discipline specificity of the proposed altmetrics system for evaluating Chinese scholars.

The winners of the Major Program of the National Natural Science Foundation of China (MP-NSFC) and the winners of the Major Program of the National Social Science Foundation of China (MP-SSFC) in 2013-2017 are chosen as the samples of the empirical study. After screening, 567 valid data are collected, including 162 winners of MP-NSFC and 405 winners of MP-SSFC. The time of the collection is from Nov. 8, 2018 to Jan. 8, 2019. We set the subject code Sub of MP-NSFC winners to be 1 and that of MP-SSFC winners to be 0.

Results

Empirical study

Through the PCA, we find that Wb_M and BD_S are equally important in two PCs simultaneously, indicating that they have little significance in the construction of the altmetrics system and should be deleted. The remaining 14 indicators are introduced into the PCA, and rotated by the maximum variance method. Then five PCs are extracted to represent these indicators and it seems that these PCs reflect five dimensions (Table 3): namely, Academic-blog attention, Social-media attention, Public-media attention, News attention, and Online-library attention. The total variance explained by the five PCs is 86%, and their rates of variance contribution are 37%, 18%, 14%, 10%, and 7%, respectively. The minimum weight of Online-library attention displays that altmetrics and traditional bibliometrics may evaluate different dimensions of scholars, which is in line with the general understanding that altmetrics represents social influence rather than academic influence of individuals (Mohammadi & Thelwall, 2014).

Table 3. Matrix of rotated principle components

	PC$_1$	**PC$_2$**	**PC$_3$**	**PC$_4$**	**PC$_5$**
Snet_C	**0.93**	0.22	0.04	0.05	0.05
Snet_V	**0.92**	0.05	0.04	0.10	0.15
Snet_M	**0.90**	0.02	0.04	0.07	0.17
Snet_R	**0.90**	0.27	0.02	0.02	0.06
BD_V	0.19	**0.92**	0.05	0.05	0.20
WeCh_M	0.22	**0.86**	0.08	0.15	0.27
BD_L	0.08	**0.84**	0.05	0.15	0.05
Wb_R	0.02	0.02	**0.97**	0.03	0.06
Wb_C	0.02	0.05	**0.96**	0.01	0.05
Wb_L	0.06	0.09	**0.73**	0.24	0.00
News_Rep	0.03	0.08	0.10	**0.94**	0.09
News_All	0.16	0.23	0.17	**0.87**	0.17
CNKI_Dl	0.07	0.35	0.02	0.10	**0.86**
CNKI_P	0.28	0.12	0.09	0.18	**0.86**

Subject specificity test

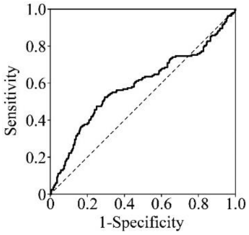


Fig 1. ROC curve when the default of Sub = 1

Setting the altmetrics score S as test variable and the subject code Sub as state variable (let default of Sub = 1), we plot the ROC curve (Fig 1.), in which the area under the curve is 0.588 ($p < 0.001$). It means that the altmetrics score we obtained did not show obvious specificity in the Social Sciences and the Natural Sciences.

Correlation test

We further explore the correlation between the altmetrics scores and the traditional measurements by choosing senior scholars in the field of library and information sciences who won the SSFC or the NSFC from 2013 to 2017 as samples. CNKI and SCI-E are retrieved to collect the data of articles and proceeding papers published by the scholars, from Feb. 21 to Feb. 28, 2019. Three commonly used bibliometric indicators, total number of articles (NA), total number of citations (NC), and h-index are calculated. After data processing, 78 valid data are obtained. Spearman correlation is used to test the relationship among the altmetrics score, NA, NC, and h-index. Results display that the altmetrics score has a positive correlation with the three traditional bibliometric indicators ($p < 0.01$), with a correlation coefficient about 0.67-0.71, implying the rationality of the score (Bornmann, 2014).

Discussion and conclusions

In this study, we introduce a range of author-level altmetrics indicators for Chinese scholars on the basis of those used in the West that are not widely accessed in China. This paper provides a reference not only for promoting the evaluation system of scholars' influence, but for researches on altmetrics system in other countries or other academic entities. While it is necessary to expand more source platforms and altmetrics indicators suitable for the evaluation of scholars, and all kinds of platforms should be encouraged to provide open-access data actively to make relative work more convenient.

Acknowledgments

We would like to thank the ISSI organisers and reviewers for their valuable opinions on the improvement of our work.

References

Priem, J. (2010). I like the term #Altmetrics#. Retrieved from twitter: https://twitter.com/#!/jasonpriem/status/25844968813

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, 8(4), 895-903.

Mohammadi, E., & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology*, 65(8), 1627-1638.

Publishing contribution of editorial board members to bibliometric indices of Library and Information Science journals

Nikolay A. Mazov[1,2] and Vadim N. Gureyev[1,2]

{MazovNA, GureyevVN}@ipgg.sbras.ru
[1]State Public Scientific Technological Library, Siberian Branch of the Russian Academy of Sciences,
Voskhod 15, Novosibirsk (Russia)

[2]Trofimuk Institute of Petroleum Geology and Geophysics, Siberian Branch of the Russian Academy of Sciences, Koptyug ave. 3, Novosibirsk (Russia)

Introduction

Editorial board members (EBMs) of academic journals are believed to be the most reputable researchers authorized to provide rigorous peer-review processes, maintain ethical principles, guarantee high quality of published materials, and consequently facilitate the advance of science. Considering a sufficient number of studies on the composition, efficiency, and functioning of EBMs, a relatively recent trend evaluating an impact of EBMs on journal bibliometric indices seems to be studied to a far less degree. EBMs may increase bibliometric indices of parent journals via publication activity including publishing in parent journal and citing parent journal in their papers in that journal (self-citations) and other sources ('hidden' self-citations). Direct publication contribution of EBMs to parent journal as compared to common authors was firstly studied by Campanario (1996). Another study (Campanario, González & Rodríguez, 2006) analyzed EBMs contribution to impact factor (IF) detecting the percentage of citations made by EBMs of the analyzed journal. This paper aims at studying the publishing contribution of EBMs to the ranks of Russian LIS journals.

Data and methods

The sample includes 22 leading LIS journals published in Russia. The sample was organized as described in (Mazov & Gureyev, 2019) and divided into three groups according to 2-year IF based on the Russian Science Citation Index (RSCI) as of 2018. The distribution revealed that the low-tiered group (IF 0.4 – 0.799) includes library journals or serials with a long history: *Automatic Documentation and Mathematical Linguistics*; *Automation and Remote Control*; *Bibliography*; *Bibliosphere*; *Information Security Problems. Computer Systems*; *Russian Journal of Library Science*; *Software Engineering*. The middle-tiered group (IF 0.8 – 0.999) includes respectable journals both on library and information science topics with a long history: *Information Resources of Russia*; *Information Society*; *Journal of Information Technologies and Computing Systems*; *Proceedings of Voronezh State University*; *Scientific and Technical Information Processing*; *Scientific and Technical Libraries*; *Systems and Means of Informatics*; *Vestnik NSU. Series: Information Technologies*; *Vestnik of Moscow City University. Series: Informatics and Informatization of Education*. The top-tiered group (IF > 1) comprises journals devoted to information science issues and is characterized by a rather recent foundation: *Business Informatics*; *Computational Technologies*; *Informatics and Applications*; *Informatics and Education*; *Journal of Applied Informatics*; *Ontology of Designing*. We analyzed papers published in 2016–2017 and citations made in 2018.

Results and discussion

Contribution of EBMs by scholarly output

EBMs papers are mainly prepared at a high level and accrue more citations. However, sometimes one may observe the opposite trend of a lower quality preparation of manuscripts by EBMs relying on superficial refereeing due to their official power (Schiermeier, 2008). Thus, it can be concluded that in the former case EBMs papers would gather a high number of citations and increase journal IF, while in the latter case they would adversely affect IF.

We failed to reveal significant dependence between IF and a share of papers by EBMs in a parent journal from the total number of journal papers: all three groups comprised approximately 16% of EBMs papers (Figure 1). Considerable differences were detected between a share of EBMs papers in a parent journal from a total number of EBMs papers in all sources. EBMs of top-tiered journals are published significantly more rarely in parent journals as compared to EBMs of middle- and low-tiered journals. When considering all journals, a share of papers in parent journals from the total number of papers in all sources was 7.6% on average.

In all three groups, we found a relatively uniform distribution of EBMs papers and usual authors, as well as a low share of papers in a parent journal as compared to the total number of papers. EBMs of top-tiered journals were found to be published more frequently in other sources rather than in a parent journal. We failed to detect any relationship between

IF and a share of papers by EBMs as it is, i.e., independently of citations.

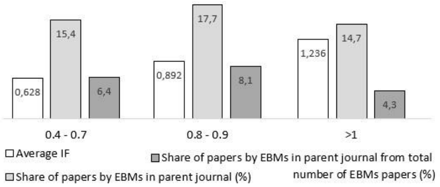


Figure 1. Relationship between IF and a share of EBMs papers.

Contribution of EBMs by citations

Citation contribution of EBMs to journal rank is twofold: either EBMs papers in parent journal can be cited or EBMs themselves can cite a parent journal from other sources ('hidden' self-citations). Figure 2 depicts both types of contributions.

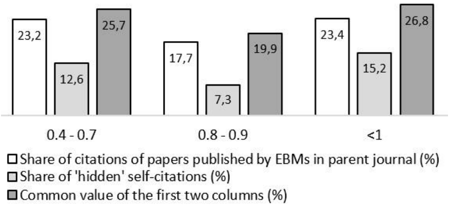


Figure 2. Relationship between IF and a share of citations to EBMs papers and/or a share of citations to parent journals made by EBMs.

Across all Russian LIS journals, 21.1% of citations belong to EBMs papers, and 11.5% of citations are made by EBMs from other sources. The common value of these two indices excluding overlapping is 23.9%. Interestingly, middle-tiered journals demonstrate lower indices as compared to top-tiered and low-tiered journals.

Dependence between IF and EBMs publication activity

To detect a direct effect of EBMs on IF, we (a) removed EBMs papers from the denominator of IF formula, (b) removed citations to EBMs papers from the numerator, and (c) removed citations made by EBMs from other sources to parent journal from the numerator of IF formula (Figure 3).

We detected a significant impact of EBMs in top-tiered and a slightly lower effect in low-tiered journals, while the middle-tiered group demonstrates a zero or reverse trend since the absence of EBMs papers in a parent journal paradoxically would have resulted in slightly higher journal ranks.

Finally, we counted the overall share of EBMs contribution according to three analyzed characteristics, i.e., EBMs publications in a parent journal, citations to EBMs publications in a parent journal, citations to a parent journal made by EBMs from other sources: 10.67% in low-tiered journals, 0.78% in middle-tiered journals, and 14.16% in top-tiered journals.

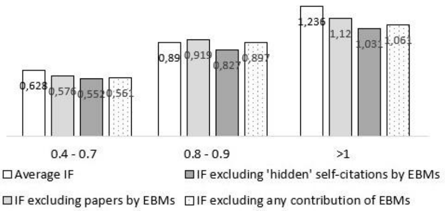


Figure 3. IF values including and excluding EBMs publishing contribution.

Middle-tiered journals stand alone because of the lack of EBMs impact on journal rank. In some journals, we revealed a negative effect (up to 18.9%) of EBMs papers on journal indices. On the contrary, some journals from top-tiered and low-tiered groups revealed a significant positive impact from EBMs publication activity on their rank (up to 34.1%). Interestingly, EBMs contribution may affect journal IF to such a great degree that some journals from our sample fell in a different subsample after the exclusion of EBMs papers and citations.

Conclusions

Analyses of the scholarly output of EBMs in Russian LIS journals enabled us to reveal some patterns in publishing contribution of EBMs to journal rank, which enhance knowledge on editorial board functionality in LIS, as well as can be used by editors-in-chief in optimizing editorial policy or changing editorial board composition.

Acknowledgments

This study was funded by the Russian Foundation for Basic Research, project no. 19-011-00534.

References

Campanario, J.M. (1996). The competition for journal space among referees, editors, and other authors and its influence on journals' impact factors. *Journal of the American Society for Information Science*, 47(3), 184-192.

Campanario, J.M., González, L. & Rodríguez, C. (2006). Structure of the impact factor of academic journals in the field of Education and Educational Psychology: Citations from editorial board members. *Scientometrics*, 69(1), 37-56.

Mazov, N.A., Gureyev, V.N. (2019). The state of Russian library and information sciences from the perspective of academic journals. *Bibliosphere*, 3, 56-70.

Schiermeier, Q. (2008). Self-publishing editor set to retire. *Nature*, 456(7221), 432.

Automatic Recognition and Classification of Future Work Sentences in Academic Articles in a Specific Domain

Wenke Hao, Yuchen Qian, Zhicheng Li, Yuzhuo Wang and Chengzhi Zhang

{haowk, qianyc, lizhicheng, wangyz, zhangcz}@njust.edu.cn
Department of Information Management, Nanjing University of Science and Technology, Nanjing (China)

Introduction

Nowadays, various digital technologies promote the publication, diffusion, acquisition, and utilization of academic literature. Therefore, researchers, mainly those who are beginners in scientific research, face rapid changes in research dynamics, and it is challenging for scholars to predict future research topics. The research paradigm characterized by 'data-driven' is emerging in the topic detection and tracking. Machine Learning (ML) and Natural Language Processing (NLP) technologies can deeply mine the rules and knowledge hidden in the data and reduce professional requirements for analysts.

Compared with the bibliographic information, full-text of academic papers contains micro-semantic details such as the chapter structure, scientific entities, and argumentation logic, revealing authors' research achievements and innovative knowledge comprehensively. The 'future work' is a pivotal part of an academic paper, where the authors make suggestions for future research, point out what research is underway, or discuss the potential direction of the whole field. Excellent future works should build on existing research and inspire new questions. Taking the field of NLP as a case, this paper extracts Future Work Sentences (FWS) of articles, classifies them into different types, and then discusses the trends of future research in NLP field. This work generates fresh insight into the analysis and mining of FWS from the large-scale full-text corpus, providing data and method support for exploring future research trends in a specific domain.

Related work

Recently, Hao et al. (2020) constructed a corpus, including future work sentences manually extracted from ACL papers published during 1990-2015, and constructed a classification system in accordance with grounded theory. Besides, most prior work is based on rules to identify future work sentences (Hu & Wan, 2015). However, manually established rules can hardly cover all the language features of future work sentences and are susceptible to subjective factors of experts. Therefore, we adopt machine learning and deep learning methods to accomplish automatic recognition and classification of future work sentences.

Method

Dataset

We download 11,952 conference papers of ACL, EMNLP, and NAACL during 2000-2019 from ACL Anthology (https://www.aclweb.org/anthology/). The manually annotated dataset consists of two parts: first, FWS of ACL during 2000-2015 are obtained from the *ACL FWS-RC* corpus (Hao et al., 2020). Second, we extract chapters related to future work in the other papers by manual reading. Subsequently, we annotate the FWS in the extracted chapters of EMNLP and NAACL according to the existing label specification and classification system. Cohen's kappa coefficient (Cohen, 1960) is employed to measure the reliability of the labels and achieves over 0.75. After deleting the connective sentences that do not contain actual meaning and splitting the FWS belonging to multiple types, we get the number of sentences in each type: *method* (4554), *resources* (987), *evaluation* (963), *application* (1150), *problem* (1043), and *other* (402). The information of annotated dataset is shown in Table 1.

Table 1. Statistic information of Dataset.

Type	#
Papers	9,508
Chapters	9,635
Sentences	62,312
Future work sentences	10,622

Model selection

For automatic recognition of FWS, we use four traditional ML models, including Naive Bayesian, Logistic Regression, Support Vector Machine and Random Forest, combining with three feature selection methods: filter, embedded, and wrapper. Then ten-fold cross-validation is performed.

For automatic classification of FWS, we use the BERT pre-training model to represent the input data's features and input the acquired feature vectors into the SoftMax layer for classification. The BERT is a model with 12 layers of Transformer, 768 hidden units, and 12 self-attention heads trained on the English dataset (https://github.com/google-research/bert). Also, we select TextCNN and BiLSTM for comparison and use Word2vec in the word embedding layer. Then the training set, verification set, and test set are divided into 8:1:1. We evaluate the model performance by Precision, Recall, and F_1.

Results

Result of Automatic FWS recognition

The experiment result shows that the Naive Bayesian based on the embedded feature selection method achieves the best F_1 score in the automatic FWS recognition. The result is shown in Table 2.

Table 2. Performance of FWS recognition.

Model	**Precision**	**Recall**	**F_1**
LR	92.46%	81.05%	85.35%
SVM	92.39%	86.87%	89.31%
RF	92.58%	92.63%	92.08%
NB	91.22%	97.58%	93.95%

Then we use this model to recognize FWS from extracted chapters of ACL between 2016 and 2019. 1430 sentences are identified and we add them to our dataset. Figure 1 shows the changes in the number of FWS. Because the amount of papers per year is not stable, the ratio of FWS is used as an indicator, which is the share of papers that contain FWS.

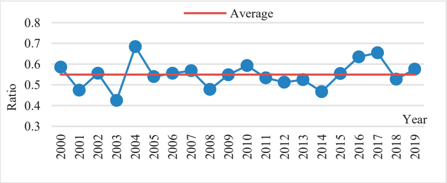


Figure 1. The ratio of FWS from 2000 to 2019.

As shown in Figure 1, the ratio of FWS in 20 years fluctuates to a small extent, with an average value of 0.55. We believe that the different types of papers influence the results, because the average value of FWS share reaches 0.58 when only long papers are considered, while the average value decreases after adding other types of papers.

Result of Automatic FWS classification

The experiment result shows that Bert performs best in the automatic FWS classification, and its performance in the test set is shown in Table 3.

Table 3. Performance of FWS classification.

Type	**Precision**	**Recall**	**F_1**
Method	89.98%	92.66%	91.30%
Resources	87.83%	82.11%	84.87%
Evaluation	91.86%	66.95%	77.45%
Application	81.32%	78.12%	80.00%
Problem	67.57%	96.15%	79.37%
Other	89.92%	63.64%	74.53%
Weighted Avg	**87.20%**	**86.02%**	**85.91%**

Similarly, we use this model to classify FWS without category labels extracted in the previous section. To enhance the result's reliability, we check the classified FWS by sampling according to the year. With a sampling rate of 10%, 22 misclassified sentences are corrected from 143 sampled sentences. Thus, we obtain a complete dataset of FWS with type labels from 2000 to 2019.

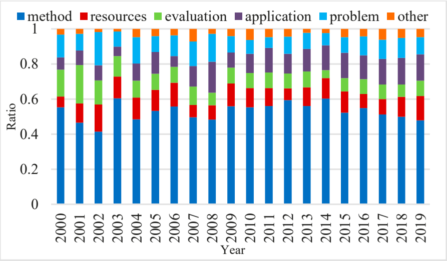


Figure 2. Ratio of FWS types from 2000 to 2019.

The evolution of the FWS types is presented in Figure 2. The ratio of *method* type each year has apparent advantages, and the remaining five types with the year present the small-scale volatility. A change from 2016 to 2019 is the dependence on data appears to increase because the ratio of *method* type is reducing while the ratio of *resources* type is rising, but it is still small compared with *method* type. Based on the result, we can infer that the future research of most papers on NLP should still focus on methods.

Conclusion and future work

This paper constructs two models to recognize and classify FWS automatically. In the future, we will apply the models to other conference papers in ACL Anthology and make the corpus publicly available to the community. In addition, we want to introduce unsupervised methods and transfer learning to reduce manual annotation and explore other areas.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 72074113).

References

Hao, W., Li, Z., Qian, Y., Wang, Y. & Zhang, C. (2020). The ACL FWS-RC: A Dataset for Recognition and Classification of Sentence about Future Works. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (pp. 261-269).

Hu, Y. & Wan, X. (2015). Mining and Analyzing the Future Works in Scientific Articles. arXiv preprint arXiv:1507.02140.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational & Psychological Measurement*, 20(1), 37-46.

International Cooperation in Times of Crises – Differences in Cooperation patterns between COVID-19 and non-COVID-19 Trials

Gerrit Hirschfeld[1] and Christian Thiele[2]

[1] *gerrit.hirschfeld@fh-bielefeld.de*
University of Applied Sciences Bielefeld, Faculty of Business, CareTech OWL Center for Health, Welfare and Technology, Interaktion 1, 33603 Bielefeld (Germany)

[2] *christian.thiele@fh-bielefeld.de*
University of Applied Sciences Bielefeld, Faculty of Business; Interaktion 1, 33603 Bielefeld (Germany)

Introduction

Long before the present pandemic took hold of the world, it has been clear that scientific Internationalization is an important aspect to how science is being conducted. Accordingly, there is a vast literature in scientometrics using bibliometric analysis to investigate networks of international cooperation in science (Glänzel & Schubert, 2004; Luukkonen et al., 1993). One previous study by Fry and colleagues (2020) investigated collaboration patterns in COVID-19 research by scrutinizing publications and pre-prints. Importantly they found that at least at the beginning of the pandemic, the earliest research collaborations tended to include fewer participating countries and also fewer emerging countries than pre-pandemic research on corona-viruses. The aim of the present study is to replicate and update these analysis using a data base of clinical registrations as a data-source.

In order to gain a picture of the research output that is as comprehensive as possible, it is often not enough to scrutinize published articles in scientific journals and books. Instead, in different fields dissertations (Andersen & Hammarfelt, 2011), and patents (Narin, 1995) are routinely used to measure scientific activities. In the domain of medicine a specific data source is available; clinical trial registers (CTRs). CTRs are data-source that are only rarely used as a data-source for scientometric analysis (for an exception see: Thelwall & Kousha, 2016). In 2008 the seventh Declaration of Helsinki established that trials involving humans have to be registered in CTRs. As a result CTRs contain information on all studies that are conducted, avoiding the otherwise pervasive problem of publication-bias, i.e. the fact that a large proportion of medical studies that were conducted are never published in scientific journals. Furthermore, since registration should happen before the first participant is enrolled, CTRs do contain information on studies that are not yet finished or have not even started.

Using CTRs as data source we want to test whether - and to what degree - international cooperation on coronavirus-research changed during the COVID-19 pandemic. And how this is influenced by pandemic-dynamics (incidents, deaths, lockdown).

Methods

Clinical Trial Registrations were accessed via the AACT-Database. This database provides monthly snapshots of the ClinicalTrials.gov registry and stores it as a relational database. Of this we used information in the “conditions” and “countries” table to identify link information.
Specifically, we compared two data sets. The first “NON-COVID” data set included interventional studies on viruses in general registered after 1.1.2010. The second “COVID” data set included interventional studies on “COVID”, “SARS”, and “CORONA” registered after 1.1.2020.

Results

Overall, we identified 1.158 NON-COVID trials and 1.955 COVID trials in the data base. As can be seen in table 1 the COVID trials had overall much fewer participating countries than NON-COVID trials.

Table 1. Number of countries involved in the NON-CORONA and CORONA dataset.

Countries	NON-COVID	COVID
1	895 (77%)	1813 (93%)
2	75 (6%)	70 (4%)
3	35 (3%)	13 (1%)
4	18 (2%)	13 (1%)
5	16 (1%)	12 (1%)
More than 5	119 (10%)	34 (2%)

Secondly, we inspected the different cooperation-networks. For this we generated two adjacency matrices; one for the NON-COVID trials and one for the COVID trials. We used scaled heat-maps to visualize these different adjacency matrices and their differences. Thirdly, we used a linear model to predict changes in cooperation intensity between two countries by COVID-19 incident-rate, mortality-rate, and differences in lockdown-start dates in the first half of 2020.

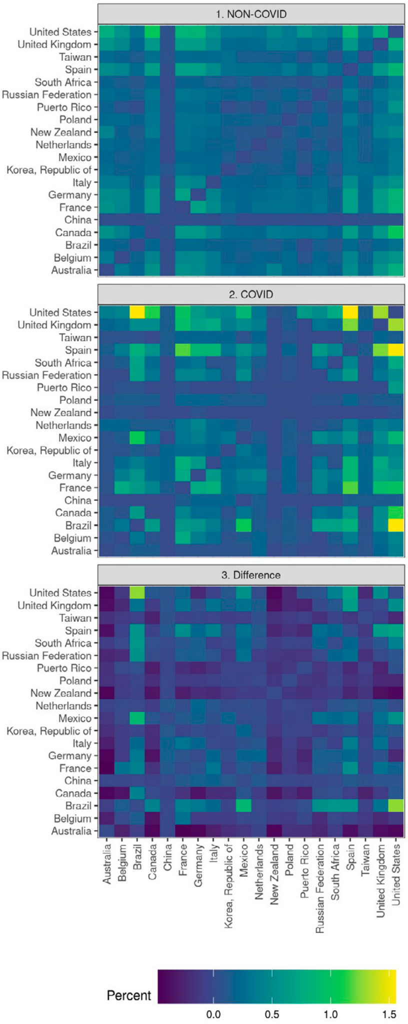


Figure 1. Cooperation networks for NON-COVID (top), COVID (middle) trials and their differences (bottom)

As can be seen in figure 1 there cooperation-pattern of NON-COVID and COVID were different in some important regards. Overall, the collaboration between the 20 countries was much more evenly distributed in NON-COVID trials compared to COVID trials. Looking at the COVID trials in particular, the many trials that were conducted in both the US, Brazil, Spain and some degree Mexico seem to stand out. At the same time, Australia and New-Zealand are less involved in COVID than in NON-COVID trials.

The linear model showed a small albeit significant association between intensified cooperation and the number of fatalities.

Discussion

The aim of the present studies was to compare the internationalization of medical research on COVID-19 to "non-COVID-19" research using CTR as data source. With regard to the overall number of involved nations, were able to replicate results based on bibliographic analysis (Fry et al., 2020) showing that COVID-19 trials conducted by smaller teams. The pattern of collaboration that emerges from our analysis of trials is however different to the pattern emerging from the bibliographic analysis. First, China is not well represented in this analysis, which may be due to the fact that there are alternative CTRs, which could be used to register trials.

Secondly, our analysis shows that emerging countries such as Brazil and Mexico are more tightly integrated into COVID trials. This may be driven by disease dynamics in the sense that countries with higher fatality-rates also increased their cooperation more strongly. Only time will tell, whether this involvement will also translate into authorship of shared publications (Glänzel & Schubert, 2004; Luukkonen et al., 1993) and further research opportunities for emerging countries.

Acknowledgments

This study was funded by the German Federal Ministry of Education and Research (BMBF #01PU17010).

References

Andersen, J. P., & Hammarfelt, B. (2011). Price revisited: On the growth of dissertations in eight research fields. *Scientometrics*, *88*(2), 371–383. https://doi.org/10.1007/s11192-011-0408-8

Fry, C. V., Cai, X., Zhang, Y., & Wagner, C. S. (2020). Consolidation in a crisis: Patterns of international collaboration in early COVID-19 research. *PloS one*, *15*(7), e0236307.

Glänzel, W., & Schubert, A. (2004). Analysing scientific networks through co-authorship. In *Handbook of quantitative science and technology research* (S. 257–276). Springer.

Luukkonen, T., Tijssen, R., Persson, O., & Sivertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics*, *28*(1), 15–36.

Narin, F. (1995). Patents as indicators for the evaluation of industrial research output. *Scientometrics*, *34*(3), 489–496. https://doi.org/10.1007/BF02018015

Thelwall, M., & Kousha, K. (2016). Are citations from clinical trials evidence of higher impact research? An analysis of ClinicalTrials. gov. *Scientometrics*, *109*(2), 1341–1351.

Changes in the Funding Composition of Czech Science

Radim Hladík

hladik@flu.cas.cz
Institute of Philosophy of the Czech Academy of Sciences, Jilská 1, 110 00 Praha (Czech Republic)

The analysis tracks funding acknowledgements of publications published by Czech scientists and reported in the Information Register of R&D results (RIV). As a national database used in authoritative research evaluations, RIV has the major advantage of exhaustive coverage of publication results in comparison with selective commercial indexing services (Sīle et al., 2018). A particular focus is on the Czech Science Foundation (CSF), the major provider of project-based funding for basic research in the Czech Republic. The breakdown of aggregated statistics into six science domains subsequently serves as an indicator of their dependence on different funding sources. The results show differentiated funding composition across scientific fields as well as a shift in the funding policies towards government-controlled sources.

The unit of analysis are the RIV records of financial resources that supported each publication. The analysis considers all publication results (articles, books, book chapters, and contributions in indexed conference proceedings) in the category of basic research, a total of 840149 publications with 1227533 acknowledgments from 2000 to 2019.

It is worth noting that RIV does not systematically record research support from foreign agencies, with some exception for EU programs. The perspective of the dataset is therefore strictly national. Furthermore, if we compare a RIV funding acknowledgment of an internationally co-authored publication with the funding data in, e.g., the Web of Science database, we typically find project acknowledgements of multiple authors. RIV records only the funding of authors based in Czech institutions. On the other hand, the Web of Science would not index non-project types of funding that make up research organizations' budgets, whereas RIV does.

Each record is accompanied by information about the scientific field according to the FORD classification. Older entries in RIV used a bespoke classificatory system, for which matching of ambiguous categories had to be resolved manually and could marginally affect the analysis.

The figure 1 shows the overall role of project-based funding in the Czech science in the structure of various financial resources declared by research organizations. The level of project funding has remained surprisingly stable and, for the past two decades, it has hovered at around 40%. In contrast, there have been substantial shifts funding policy in the composition of the remaining 60% of acknowledgements (cf. Good et al., 2015). Forward-looking institutional funding based on "research plans" has been gradually replaced by institutional "support for the development of research organizations" distributed according to post-hoc research evaluation. Universities have also received earmarked funds for "specific higher-education research," which requires student participation. Other types of financial resources play a smaller role.

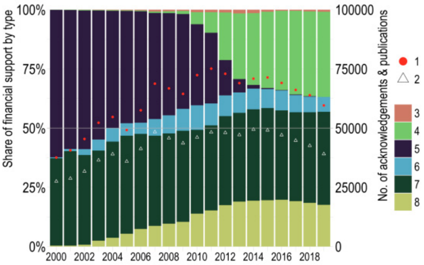


Fig. 1. Financial sources in the publication acknowledgements.

1 – No. acknowledgements. 2 – No. publications. 3 – Other public or private sources. 4 – Institutional support for development. 5 – Institutional support for research plans. 6 – EU Programs. 7 – Projects. 8 – Specific higher-education research.

In the next step of the analysis, we focus only on project dedications, i.e., roughly less than 40% of all acknowledgements. While the overall share of project resources has remained stable, there have been more significant changes within the project segment itself.

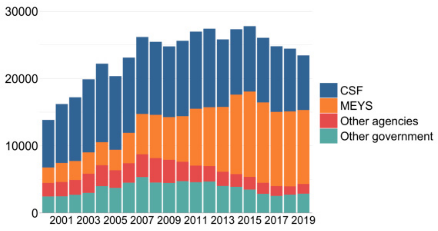


Fig. 2. Project acknowledgments by providers.

As the figure 2 shows, the role of the CSF has weakened over the years. While until 2005 inclusive, CSF could claim more than half of all project dedications, by 2019 its share was just over one third. Smaller funding agencies, regional providers and even governmental sources have likewise seen a comparative decline in the received acknowledgements. The only provider to increase its share has been the Ministry of Education, Youth, and Sport (MEYS) which has been charged with national redistribution of EU funds. By 2019 MEYS has been receiving almost half of all funding acknowledgements.

The section B of the figure 3 shows a mix of project-based funding providers by science domain. It breaks down the aggregated statistics discussed above by major scientific fields in the FORD classification. The simple count of funding acknowledgement implies a degree of standardization, whereby disciplines are compared by the number of publications rather than by the amounts of money required to carry out the research underlying those publications (cf. section A of the figure 3).

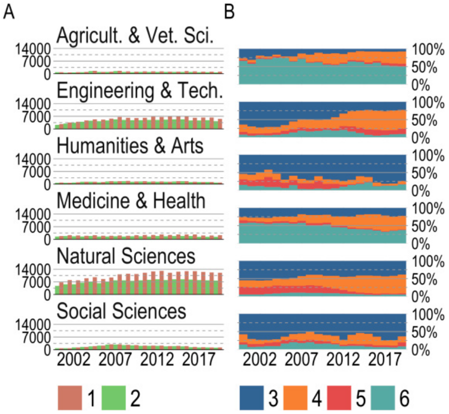


Fig. 3. Project acknowledgements by fields.

1 – No. acknowledgements. 2 – No. publications. 3 – CSF. 4 – MEYS. 5 – Other agencies. 6 – Other government.

The picture reveals three main patterns. 1) Agricultural and medical disciplines are supported primarily by governmental bodies. They too have experienced a greater influence of MEYS projects. 2) Project resources from MEYS dominate in the natural sciences and in engineering, although CSF still holds an important position in the natural sciences. 3) Humanities and social sciences largely depend on CSF for project funding. Their opportunities to obtain project funding from other providers are severely limited.

The figure 4 applies disciplinary resolution again to all sources of funding. Project acknowledgements account for only about 1/5 of all dedications in the humanities and social sciences and for about a half of the acknowledgements in the agricultural, technical and natural sciences. At this level, fields cluster differently than they did in the exclusively project segment. Medical, social and the humanities research report institutional support in the majority of dedications. Although Social Sciences and the Humanities depend most heavily on the CSF for project funding, project-based funding is generally less important in their overall publication output. Support from CSF is of greater importance in the Natural Sciences. This is illustrated by the fact that in 2015-2019, 14% of all dedications (i.e., including non-project sources) were attributed to CSF from the Humanities, compared to 20% in the Natural Sciences.

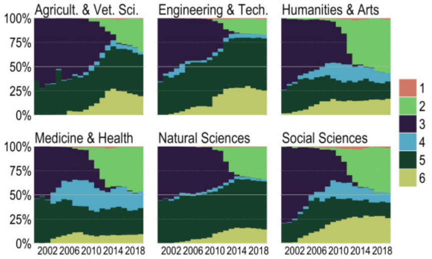


Fig. 4. Financial sources by field.

1 – Other public or private sources. 2 – Institutional support for development. 3 – Institutional support for research plans. 4 – EU Programs. 5 – Projects. 6 – Specific higher-education research.

The analysis did not consider the actual amounts of financial support. It also could not account for the influence of foreign funding providers. Despite these limitations, we can conclude that the strategic role of the national grant agency varies significantly across different scientific fields. Scientific funding agencies are a special space on the frontier of science policy and science itself (Braun, 1998). Thanks to expert panels and an emphasis on peer-review, the voice of the scientific community resonates in them. The state’s growing involvement in the project-based funding of science at the expense of the main national grant agency therefore signals an important shift in the relative autonomy of science funding mechanisms in the Czech Republic.

Acknowledgments

This research was supported by The Czech Science Foundation project no. GJ20-01752Y, “Funded and Unfunded Research in the Czech Republic”.

References

Braun, D. (1998). The role of funding agencies in the cognitive development of science. *Research Policy*, 27(8), 807-821.

Good, B., Vermeulen, N., Tiefenthaler, B. & Arnold, E. (2015). Counting quality? The Czech performance-based research funding system. *Research Evaluation*, 24(2), 91-105

Sīle, L. et al. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. *Research Evaluation*, 27(4), 310-322.

Using an Entitymetric Approach to Trace Novel Chemical Compounds for Research Impact

David E. Hubbard

hubbardd@library.tamu.edu
Texas A&M University, University Libraries, 5000 TAMU,
College Station, Texas (USA)

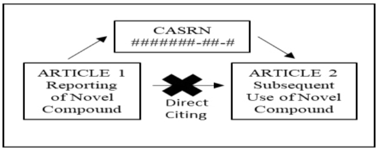
Introduction

This study presents an entitymetric approach (Ding et. al, 2013) to trace the use of novel chemical compounds through the literature to demonstrate additional research impact beyond traditional citations. The development of novel chemical compounds is the intellectual embodiment of the research of many organic chemists, and therefore worth tracing to capture a more complete picture of their impact. Citation-based approaches certainly capture some of the impact, though overlooks other aspects of impact. While this study focuses on chemical compounds, the concept could be more broadly extended to other material objects that contribute to advancing research (e.g., tools, software, etc.). The main purpose of this study is to outline the approach using a small set of journal articles from an individual researcher to illustrate the idea and process, as well as to determine if there are indeed occurrences where the novel chemical compounds are utilized and the original synthesis or isolation are not cited in the traditional manner.

Background

SciFinder is the most comprehensive bibliographic database in chemistry. In addition to abstracting and indexing the literature of chemistry and related disciplines, it assigns a unique Chemical Abstracts Services Registry Number (CASRN) to each new chemical compound reported in the literature that SciFinder indexes. CASRNs are themselves indexed, as well as linked to all future publications that use that compound. This provides a means to trace the first published report, which is often a synthesis or isolation of that compound, to any subsequent uses described in publications. Of particular interest are subsequent articles that use a novel compound in a substantive way (e.g., to facilitate a chemical synthesis or application), but do not cite the original article where the compound was first reported (Figure 1).

Figure 1. Linkage through an entity (CASRN)



There may be legitimate reasons for not citing the original article reporting the first novel synthesis or isolation. For example, the researchers may have synthesized the compound themselves through another synthetic process or purchased it. Furthermore, it is not a norm for chemists to cite articles associated with the first reported synthesis of every chemical utilized in experimental methods. So if such instances exist, this would represent a source of unrecognized research impact that can be traced through the scientific literature.

Literature review

There are no known studies that trace chemicals or CASRNs through the literature as described in this study. There are a few studies that have used the CASRNs to either select publications for a bibliometric study (Bornmann & Marx, 2013) or compound-based bibliometric studies that map and identify research gaps (Barth & Marx, 2012; Tomaszewski, 2019); however, the compounds in those studies were not traced as described in this study. A similar study, at least conceptually, is that of Ding et al. (2013) where an entity-entity citation network for the drug Metformin was used to identify connections with diseases, drugs, and genes. While the present study uses the concept of an entity (novel chemical compounds), its focus is on locating subsequent use of novel compounds where the original article reporting its first synthesis or isolation was not cited.

Methods

The publications of an organic chemist were identified in SciFinder for the following years: 2000, 2005, 2010, and 2015. The chemist's publications were refined to peer-reviewed journals reporting original research, so review articles and other types of publications indexed in SciFinder were excluded. The compounds, or rather the unique CASRNs associated with those compounds and designated as being "synthetic preparations" in the SciFinder record for each article were identified. Each of the hyperlinked CASRNs in SciFinder associated with those synthetic preparations in the article record were clicked resulting in all the publications mentioning that compound within SciFinder. If the oldest publication was the original peer-reviewed journal article of the organic chemist, the compound was deemed "novel" (i.e., the organic chemist was the first to report the synthesis or isolation of that compound in the literature). If the compound was novel, then all publications citing the CASRN were examined to: (1) determine how the compound was used in subsequent research, and (2) determine if the researchers cited the original article reporting on the first synthetic preparation or isolation.

Results

Table 1 summarizes the number of articles, chemical compounds, and number of novel chemical compounds by year for an individual organic chemist. It should be noted that subsequent use by the organic chemist reporting the original synthesis or isolation was excluded (i.e., "self-citation"), as were compounds associated with patent applications filed by the organic chemist since those were filed before the articles and therefore not novel. On average, 62% of the compounds synthesized in those articles were novel. Of the 230 novel compounds, 8 were utilized or mentioned in articles of other researchers (Table 2). Those uses (or roles) were characterized as being an intermediate (I), product (P), or reagent (R). These roles, and others, accompany the CASRNs listed in ScFinder records. The most interesting and potentially impactful instances are the novel compounds being used as reagents (i.e., starting material) to facilitate the synthesis of other chemical compounds; whereas the intermediate and product are reporting on its presence in the middle of a multistep synthetic reaction (intermediate) or new syntheses of the novel compound (product). The focus here was primarily on those being used as reagents, where there were five articles that utilized the novel compound as a reagent. Of those five articles, two did not cite the original article. So among the 230 novel compounds only 2 (or 0.9%) were used as reagents without citing the original article reporting the first synthetic preparation or isolation. Based on SciFinder, one compound is now commercially available (CASRN: 269749-51-1) and the other compound is not (CASRN: 871117-18-9). In the case of the former, the article does not cite an article or commercial source. In the case of the latter (CASRN: 871117-18-9), the article does not appear to contain a synthesis for that compound. Without contacting the authors of these two subsequent articles utilizing these two novel compounds, it is difficult to determine if the citations to the original synthesis were just not included (and therefore unrecognized impact) or if the two novel compounds were obtained through another means.

Table 1. Number of Articles, Compounds, and Novel Compounds

Year	Articles	Cmpds Prepared	Novel Cmpds
2000	8	126	95 (75%)
2005	5	80	65 (81%)
2010	4	123	43 (35%)
2015	3	43	27 (63%)
Total	20	372	230 (62%)

Table 2. Novel Compounds Subsequently Utilized, Roles, and Uncited

Year[a]	Novel Cmpds Utilized	Articles Utilizing Novel Cmpds and Role	Articles Utilizing Novel Cmpds and Role w/o Citing
2000	5	7 (1-I;2-P;4-R)	2 (1-I;1-R)
2005	1	1 (1-R)	1 (1-R)
2010	1	2 (2-P)	2 (2-P)
2015	1	1 (1-I)	1 (1-I)
Total	8	11	5

[a] Year only associated with the Novel Cmpds Utilized column.

Conclusion

This study demonstrated how the CASRN (an entity) can be traced through the literature, and that there were indeed occurrences where novel compounds are used by other authors without citing the original synthesis or isolation (0.9%). Based on the findings of this study, it is recommended that larger quantitative and qualitative studies are conducted to substantiate the scale of this impact and the nature of the omissions since this study was merely proof of concept. The manual methods employed in this study were very tedious and relied on a proprietary database. Therefore, it is also recommended that the process be automated to identify the uncited novel chemical compounds more efficiently, as well as explore the use of open chemical data sources containing CASRNs for increased accessibility and transparency.

References

Barth, A. & Marx, W. (2021) Stimulation of ideas through compound-based bibliometrics: counting and mapping chemical compounds for analyzing research topics in chemistry, physics, and materials science. *ChemistryOpen*, 1, 276–283.

Bornmann, L. & Marx, W. (2013). The proposal of a broadening of perspective in evaluative bibliometrics by complementing the times cited with a cited reference analysis. *Journal of Informetrics*, 7, 84-88.

Ding Y., Song M., Han J., Yu Q., Yan E., Lin, L. & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. *PLOS ONE*, 8, e71416.

Tomaszewski, R. (2019). Substance-based bibliometrics: Identifying research gaps by counting and analyzing substances. *ACS Omega*, 4, 86-94.

A Bibliometric Mapping of Research on Quality Assurance in African Higher Education: a systematic literature review (from 1999 to 2019)

Luís M. João[1, 2] and Patrício Langa[1, 2]

[1] *3957821@myuwc.ac.za and planga@uwc.ac.za*
University of the Western Cape, Robert Sobukwe Rd, Bellville, Cape Town 7535 (South Africa)

[2] *luis.m.joao@uem.ac.mz and patricio.langa@uem.mz*
Eduardo Mondlane University, Main University Campus, Julius Nyerere Avenue, Maputo 3453 (Mozambique)

Introduction

The education sector, in general, and higher education institutions (HEIs) in particular, use 'quality' as a common term, yet the notion of quality is disputed and multi-dimensional (Ewell, 2010; Zabadi, 2013). Therefore, quality revolves around a few central ideas: quality as absolute; quality as relative; quality as a process, and quality as culture (Harvey & Williams, 2010; Kleijnen et al., 2013). Furthermore, quality can be based on the ability of an institution to fulfil the requirements of the academic community.

Quality assurance (QA) is the process of establishing stakeholders' confidence that the provision of education (inputs, processes, and outcomes) will fulfil the expectations of stakeholders. From the 1990s onwards, QA has been on the contemporary higher education (HE) agenda and has been prioritised around the world (Cardoso et al., 2019; Elken & Stensaker, 2018).

The economic value of higher education has led some scholars to examine the trends in QA research in higher education (Serrano-Velarde, 2008) to understand how quality can be maintained and improved for higher education systems.

The poster reflects the research we are doing on QA of higher education in Africa. Specifically, it shows a correlation between the increase in the number of students pursuing (higher) education degrees and the rise of literature on QA. We specifically focussed on QA in African higher education because, as African based scholars, we see first-hand the advantages that QA research can bring to the improvement of higher education as whole. Through a research methodology that uses systematic literature review, we have gathered information over a 20-year period (1999-2019) from the following four sources: WoS, Scopus, AJOL, Google scholar, as way to map the emerging field of research on QA in Africa.

Research problem

For over three decades, African countries have looked at tertiary education to make a significant contribution to economic growth and competitiveness, improve the quality of educational programmes and institutions (Wangenge-Ouma & Langa, 2011; Materu, 2007). This has resulted in a growing interest among scholars on how to manage QA through research based evidence (Al Jaber, 2018; Coates, 2005; Federkeil, 2008; Harvey & Williams, 2010). The evidence-based QA management endeavours to make QA processes more intelligible and rigorous. Through research the strong convictions and somehow fragile evidence that informs policy and public debates on QA in African HE is addressed.

However, while there has been a growing stock of knowledge and research-based evidence on QA through research, monitoring, and evaluation of African higher education, QA research remains not widely accessible to policy-makers and the larger public.

The scattered nature, organisation and management of QA knowledge, as in themes and approaches, has led to uninformed policy action to improve QA practice, especially in the African higher education context.

Therefore, this study seeks to produce a synthesised portrayal themes and approaches on the basis of scientific publications on African QA.

Research question

What kind of research has been conducted on QA in African HE published from 1999 to 2019, and its implication on policy and practices in the African higher education?

Methodology

Research design

To map the QA research, and its implication on policy and practices in the African HE, this study undertakes a systematic literature review focusing on QA research in Africa. Moreover, a bibliometric analysis is used to scrutinise books, articles, and other publications on African QA field.

Sampling

In executing this study, a purposive sample search of peer-reviewed articles, books, and proceedings in high rated journals, published from 1999 to 2019 on 54 African countries, has been conducted in the main scientific databases, namely Web of Science (WoS), Scopus, Google Scholar, and African Journal Online (AJOL). These main scientific databases are suitable for in-depth evaluation of complex issues (Creswell, 2014).

Data collection instruments

The data collection process was undertaken, first, through a systematic literature review focusing on QA research in Africa, based on PRISMA Statement (Moher et al., 2009). This procedure was then followed by a Bibliometric analysis and citation mapping process.

Data presentation and analysis

The data presentation and analysis is based on a combination between bibliometric analysis that relies on citation data and the visualization technique (Steinhardt et al., 2017; Alzafari, 2017; Swain, 2014; Hallinger & Chatpinyakoop, 2019). The study follows six steps (i.e., Research objectives; Research design; Bibliometric data collection; Methodology and software; Analysis and results; and Interpretation of findings) as summary of research design (i.e., adapted from Bragazzi (2019), and Zancanaro et al. (2015)).

Preliminary results

The primary purpose of the study was to investigate the state of art of the African QA emerging field of research through a systematic literature review. Table 1 and figure 2, show the rational and structural attributes of the African QA field of research conducted through bibliometric analysis.

Table 1. The African regions and countries evolvement in QA in HE research from 1999 to 2019 and retrieved from Scopus and WoS.

ID	Region	No.Article	Citation	Country with most impact	Percentage	
					No. Article (%)	Citation (%)
1	North Africa	5	9	Egypt (3/4), Morocco, Tuni	6	3,4
2	West Africa	14	29	Ghana (5/16), Nigeria (8/1	16,8	11,1
3	Central Africa	2	4	Angola (2/4)	2,4	1,5
4	East Africa	18	88	Kenya (5/56), Ethiopia (2/2	21,7	33,6
5	**Southern Africa**	44	132	**South Africa (41/127)**, Nan	53	50,4
6	**Whole Africa**	**83**	**262**		**100**	**100**

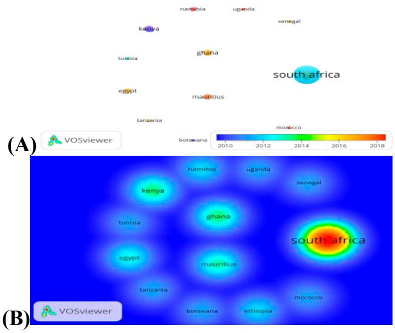


Figure 2: QA in African HE: (A) Countries' Citations Overlay visualization; (B) Countries' Citations Density visualization.

Based on table 1 and figure 2 (data from Scopus and WoS), the evolvement of African regions is visualized, especially from South Africa (SA). Therefore, figure 2A shows that, 50 articles of 83 were published between 2006 and 2015 signalling 56.9% of literature compared to all. Therefore, where 50.4% contribution were published from SA.

Conclusion

In conclusion, there is a lot of research and publication that has been done since 1999 to 2019. Therefore, the review looked at different approaches and themes to QA in African HE. Based on systematic literature review, QA research in Africa approaches to Individual based, Department based or Professional community-based at community level control, Regional level control, National level control, Continental harmonization, and international approaches.

As for the themes, we gathered themes as Definition of QA; Purpose or function of QA systems in different countries and regions; Methodology for quality assessment; Strategies to ensure quality; quality audits; Accreditation, national rankings and global rankings; and challenges, opportunities, globalization of QA.

Based on the data that we have now (table 1 & figure 2), looking into the selected research period, the preliminar results shows that papers published, from four sources, were 3 in 1999 to 503 in 2019 signalling 99.4 % of literature increase. Therefore, we can conclude that a lot of research and publication has been done since 1999 to 2019, when it comes to research on QA in African HE.

References

Al Jaber, A. O. (2018). *Toward Quality Assurance and Excellence in Higher Education*. Alsbjergvej: River Publishers.

Alzafari, K. & Ursin, J. (2019). Implementation of quality assurance standards in European higher education: Does context matter? *Quality in Higher Education,* 25(1), 58-75.

Coates, H. (2005). The value of student engagement for higher education quality assurance. *Quality in Higher Education*, 11(1), 25-36.

Cardoso, S., Rosa, M. J., Videira, P. & Amaral, A. (2019). Internal quality assurance: A new culture or added bureaucracy? *Assessment & Evaluation in Higher Education*, 44(2), 249-262.

Federkeil, G. (2008). Rankings and quality assurance in higher education. *Higher Education in Europe*, 33(2-3), 219-23.

Freeman, L. C. (2011). The development of social network analysis: With an emphasis on recent events. In Scott, J. & Carrington, P. J. (Eds.), *the SAGE Handbook of Social Network Analysis*, (pp. 26-39). California: SAGE Publications.

Harvey, L. & Williams, J. (2010). Fifteen years of quality in higher education *Quality in Higher Education*, 16(2), 81-113.

Hazelkorn, E., Coates, H. & McCormick, A. C. (Eds.). (2018). *Research Handbook on Quality, Performance and Accountability in Higher Education*. Cheltenham: Edward Elgar Publishing.

Materu, P. N. (2007). *Higher Education Quality Assurance in Sub-Saharan Africa: Status, Challenges, Opportunities, and Promising Practices*. Washington: World Bank.

Steinhardt, I., Schneijderberg, C., Götze, N., Baumann, J. & Krücken, G. (2017). Mapping the quality assurance of teaching and learning in higher education: The emergence of a specialty? *Higher Education*, 74(2), 221-237.

Stensaker, B. & Harvey, L. (Eds.). (2010). *Accountability in Higher Education: Global Perspectives on Trust and Power*. New York: Routledge.

Swain, D. K. (2014). Journal bibliometric analysis: A case study on quality assurance in education. *Indian Streams Research Journal*, 4(4) 1-15.

Measuring Altmetric Events: The need for Longer Observation Period and Article Level Computations

Mousumi Karmakar[1], Sumit Kumar Banshal[2] and Vivek Kumar Singh[3]

[1]*mousumi.kamakar10@bhu.ac.in,* [3]*vivek@bhu.ac.in*
Department of Computer Science, Banaras Hindu University, Varanasi (India)

[2]*sumitbanshal06@gmail.comt*
Department of Computer Science, South Asian University, New Delhi (India)

Introduction

The tracking and measurement of social media mentions to scholarly articles has gained a lot of attention during last few years. Some studies have exclusively focused their attention on the speed of accumulation of these altmetric events, concluding that the altmetric attention is quite quick to accrue but at the same time it may also decay quickly (Ortega, 2018; Haustein, 2019; Fang & Costas, 2020). It has been observed that most of the studies on measuring altmetrics use limited observation periods, usually ending when majority of altmetric events were seen. On the contrary, a careful look at the altmetric data for a larger observation period, shows that altmetric accumulation continues beyond the immediate-past of article publication. Further, previous studies have defined half-life value over a set of articles rather than doing an article-level delineation. This may be problematic given the fact that altmetric mentions to scientific articles vary by demography, discipline etc.

This article, therefore, attempts to demonstrate the weakness of using limited observation periods and set-level computations and emphasize the need for longer observation period and article-level computations for measuring the altmetric events. We show, through a large-sized data analysis, that the values of altmetric measures change if a longer observation period is taken. Further, it is also argued that computation of altmetric measures (such as half-life) should be done through an article-level analysis rather than computing for the whole set of articles.

Data & Methodology

The data for analysis was taken from 1,785,149 publication records for the whole world for the year 2016 as indexed in Web of Science (WoS). A total of 1,661,477 of the total records, had altmetric data (tweets) available in PlumX aggregator. PlumX gives the counts, including retweets, but more detailed information like tweet text, tweets, tweet date etc. is available only for the original tweets. Since we needed the date information for the tweets, therefore we have limited our analysis to original tweets only. Out of the total data, we selected 24,290 articles, that had a minimum of 8 original tweets. We obtained the creation date of article from the 'DOI creation date' field in Crossref. For each of the 24,290 articles, dates of all their original tweets are obtained by following the link 'Plum stable url' and scraping their original tweets. The crawling/ scraping have been performed using web-based crawler written in python programming language. To avoid overloading the server a random delay between 10 to 180 seconds has been put after each step. This way we have obtained tweet data along with the time line for the set of 24,290 articles over a period of 4.5 years.

To understand the dynamics of tweet accumulation around a scholarly article, two previously defined time-dependent variables have been examined. These time related measures, discussed and tested in recent studies, are namely, *half-life* and *velocity index*. The half-life measures the number of days to accumulate half of the tweets and can be defined by the following formula:

$$h = d(\sum tweets/2)$$

where, d is the number of days. The velocity index (VI) for any set of articles, at any time can be defined as the proportion of tweets accumulated until that time, defined as follows:

$$VI = \frac{P_i}{TP_i}$$

where, P_i = tweet accumulated until specified time, TP_i = total tweets during the observation period.

Results & Discussion

The article mainly computed two time-oriented measures, the half-life and the velocity index (VI). First, we describe how these computations are affected by taking longer observation period and thereafter the impact of article-level computations are presented.

Longer observation period

While previous studies used a shorter time span for half-life & VI computations, we wanted to demonstrate that taking a longer period is likely to

Table 1: Tweet half-life values for data from different observation periods

	180 days	360 days	540 days	720 days	900 days	1080 days	1260 days	1440 days	1620 days
Half-life	6	8	12	15	19	24	30	36	41

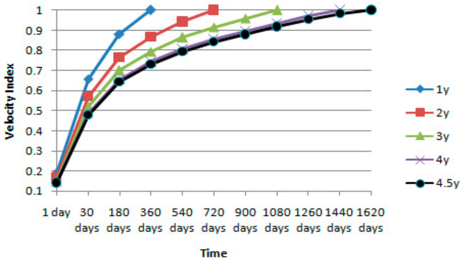


Figure 1: Variation in Velocity Index with different observation period

change these values significantly. **Table 1** presents the half-life values of the set of articles computed for different observation periods, varying by 6 months. These values are computed over the whole set. It is observed that the half-life for 6 months period is only 6 days. For 1.5 years period, it becomes 12 days, indicating that taking data for one more year, doubles the half-life value. For 2-year observation period, the half-life is 15 days. Our results show half-life of 41 days at the end of 4.5 years, which is 3 times more than that reported by Fang & Costas (2020). These values indicate that tweets keep on accumulating over a much longer observation period, unlike the commonly held belief that tweet accumulation decay quickly.

The velocity index (VI) values have also been computed at different time intervals for different observation periods, as shown in **Figure 1.** It can be observed that for a 1-year time window, 19% of total tweets accumulate during 1st day and about 88% tweets accumulate during 6-months. When the observation period is extended further, the velocity index values at the same time interval change significantly. Thus, moving from observation period of 1-year to 4.5 years, it is observed that only 62% tweets are accumulated during the 1st year i.e., 38% more tweets accrue afterwards. This is different from observation of Fang & Costas (2020), which found that 90% of the tweets accrued within the first year. Therefore, it is better to take longer observation period for computing different time related measures of altmetric events.

Article-level computations

The previous studies have computed the half-life and VI values at the level of the whole set of articles. Given that altmetric distributions are highly skewed, computing the half-life and VI values at the set level fails to capture the skewness in the distribution, rather the outliers may introduce distortions in the computations.

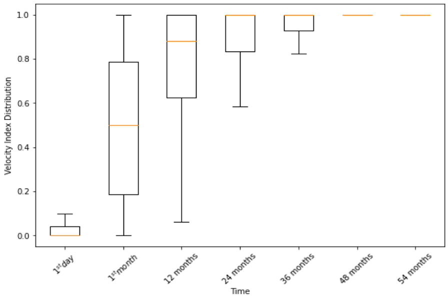


Figure 2: Variation in Velocity Index (VI) for the articles

Figure 2 shows the values of VI for the entire set of articles at different time intervals. On the first day, the VIs of all the articles is nearly 0 except for some outliers. For the 1-month period, variation in VI values is found to range between 0.2 to 0.8. Similarly, for 1-year period, the VI values for articles range from 0.6 to 1.0. Similar variations are observed for other time intervals. The computation of half-life shows a similar pattern of variations. Thus, a set level computation of half-life and VI may not be a suitable representation of the wide range of values for different articles. Further, the high variation in VI observed in the shorter observation periods, which diminish as we go for longer observation period of 4 years or more, also supports the need for longer observation periods.

Conclusion

The analytical results show that using different observation periods change the measured values of the time-related measures. The results imply that longer observation period should be used for appropriate measurement of altmetrics. Further, the use of article-level delineation for computing the measures is advocated as a more accurate method to capture the true values of such measures.

References

Fang, Z., & Costas, R. (2020). Studying the accumulation velocity of altmetric data tracked by Altmetric.com. *Scientometrics*, *123*(2), 1077–1101. https://doi.org/10.1007/s11192-020-03405-9

Haustein, S. (2019). Scholarly twitter metrics. In *Springer Handbooks* (pp. 729–760). Springer. https://doi.org/10.1007/978-3-030-02511-3_28

Ortega, J. L. (2018). The life cycle of altmetric impact: A longitudinal study of six metrics from PlumX. *Journal of Informetrics*, *12*(3), 579–589. https://doi.org/10.1016/j.joi.2018.06.001

Tutmetrics and usability: The footprint of science mapping tools in tutorial space.

Veslava Osinska[1] and Radoslaw Klimas[2]

[1]*wieo@umk.pl*
Nicolaus Copernicus University, The Institute of Information and Communication Research, Bojarskiego 1, 87-100 Toruń (Poland)

[2] *rk@doktorant.umk.pl*
Nicolaus Copernicus University, Doctoral School of Social Sciences, Bojarskiego 1, 87-100 Toruń (Poland)

Introduction

Dedicated to specific research tasks applications can vary in the scope of functions, usability degree, audience, specialization, and architecture. Crucial front-end features of software include its structure, language, examples, syntax and form and associated documentation. Video tutorials have an attractive form that can not only help in making decisions but can also educate on the various tool functions. Many well-prepared tutorials help with broadening the audience for these tools; the number of videos about any topic can be correlated with topic's popularity.

Science mapping (SM) and visualization is largely a practical field in which the accessibility of tools and proficiency in their use, as well as knowledge about bibliographic databases, statistics, and visualization techniques, determine the success of complex, multifaceted analyses. SM researchers have analyzed different sets of tools. The largest and most complete set contained 20 software examples (Chen, 2017; Bankar & Lihitkar, 2019; He et al, 2019) and Gephi (Bastian, 2009).

The authors decided to compare a set of SM tools in terms of functionality, usability, and tutorial accessibility on the web. They also strove to determine if these areas had any influence on each other. Information architecture (IA) and usability expertise review and YouTube statistics were used to examine the most significant scientometric tools since their creation.

For the current work, the authors composed a final set of SM software, the common in previous studies. The final list consisted of eight applications: Pajek (1996), CiteSpace (1999), HistCite (2004), Gephi (2009), Sci2 (2009), VOSviewer (2010), SciMAT (2016), and CitNetExplorer (2017). The period of comparison extends over 25 years, distributed across the decades.

Methods

If we study accessibility and the usage of the tools' tutorials published on YouTube or other social media, we can evaluate the popularity of a specified software tool among users (i.e., researchers). This perspective of studying the footprints of tutorials on the web has been termed *tutmetrics*. Usability and feature analysis help determine the best and most versatile ones. In the tool comparison, both the professional experience in user experience (UX) and software usability of the authors were used.

Results

Statistics of tool use

Selected tool references were examined in WoS database. A query was constructed by entering each name in the topic field and excluding the others. The Table 1 contains comparison results according: tools life length and its footprint from the WoS literature and YouTube. The usage index according to the WoS means the number of articles related to the tool, normalized by its life length. Similarly, the second usage index was calculated using the number of YouTube tutorials.

Availability of video tutorials

Data were obtained for 350 uploaded video tutorials. The total number increased from nine (in 2010) to 49 (2018) and then 84 (2019). Dynamic data of the uploaded tutorials is omitted due to publishing limitations. It should be noted that the delay time from the initial publication year of a tool to its first video tutorial varied: 0 (SciMAT), 1 year (Gephi), 3 years (Sci^2), 5 years (VOSviewer), eight years (HistCite), 13 years (Pajek), and 15 years (CiteSpace).

The titles of the video tutorials were also analyzed in terms of frequency (excluding the names of the tools and authors). The most common words in the titles were: "tutorial," "network," "analysis," and "introduction" and the phrases: "network analysis," "avaliar produção," "bibliometric analysis," "citation analysis," and "social network". The languages of the videos are shown in Figure 1.

Expert assessment of usability features

To easily operate the tool and classify it as user friendly, expert review had to be applied in two main categories: information architecture (labelling, navigation, organization, help system) and the tool's usefulness based on its features.

Table 1. Usage of the most important tools for SM since 1996

Application	Cite Space	VOS viewer	Gephi	Pajek	Hist Cite	Sci MAT	Sci^2 Tool	CitNet Explorer
Life length	21	10	11	24	16	4	11	6
Articles	824	754	324	261	151	79	37	20
Usage index in WoS	40.0	75.4	29.5	10.9	9.4	19.8	3.4	3.3
Number of video tutorials	15	40	187	45	43	9	9	2
Usage index on YT	0.7	4.0	13.4	1.8	2.7	1.1	0.8	0.0

The last ones related to the accessibility of functions, such as format compatibility, visual layout choice, clustering, labeling, filtering, statistics, and the manipulation available to the user. The full process of interaction with the software starting from installation and ending with generating visual layouts was evaluated using a 6-point scale. Because of limited space the detailed table cannot be presented. Aspects that needed time measurements were tested on the same larger dataset to standardize the conditions.

Discussion and conclusions

The top ranks in both the scientific literature and YouTube reflect the modern generation software. The ranking put the Gephi, which is highly visible in scientific databases and multimedia resources in the first place, while VOSviewer - in the second. Despite the high popularity of Pajek and CiteSpace in the literature, they were not largely represented on YouTube.

Gephi has been seen on YouTube since its initial release and has shown consistent growth. However, most of the tools had not been represented with any new videos for a few years. Gephi had more tutorials than all the other applications combined. The high number of English, Spanish, and Portuguese video tutorials suggests the dominating languages.

Expert analysis has shown that VOSviewer and Gephi have the best IA. Pajek's and HistCite's older layouts do not prevent them from being functional tools. Contextual hints and official video tutorials explaining features will also be helpful for all of the tools. Despite all the differences in visual and formal attributes, the identified tools can all be very useful for different needs.

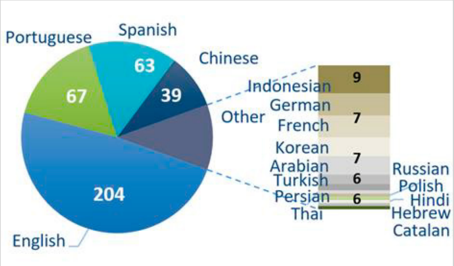


Figure 1. Frequency of the languages of the video tutorials describing SM tools.

SciMAT and CitNetExplorer, the most recent tools examined, were also scored highly in terms of usability and seem to have a focus on widely understood multidisciplinary SM.

The results suggest that there can be a correlation of tool feature usability with its presence in YouTube tutorials. The authors studied and compared these basic SM applications in terms of accessibility for learning and usability. They analyzed the accessibility of video tutorials on YouTube and application interfaces in terms of information architecture. A new term was introduced, tutmetrics, to emphasize the importance of the current study in the context of the multidisciplinarity of mapping scientific literature.

Only three of the considered tools were referenced frequently by both scientific articles and YouTube tutorials: modern generation, user-friendly software VOSviewer and Gephi and Pajek with the longest history. It is worth noting that the very popular, as well as traditional, bibliometric software CiteSpace was practically invisible on YouTube; thus, from these correlations, we cannot make definitive statements about effect-causal relationships. Does high popularity among researchers and high usability generate more tutorials, or does a large YouTube tutorial space imply the use of the tool for research? These questions suggest new possibilities for tutmetrics in future research.

References

Bankar, R., & Lihitkar, S. (2019). Science Mapping and Visualization Tools Used for Bibliometric and Scientometric Studies: A Comparative Study. *Journal of Advancements in Library Sciences*, 6(1).

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open-Source Software for Exploring and Manipulating Networks. *Proceedings of Third International AAAI Conference on Weblogs and Social Media,* 3(1).

Chen, Ch. (2017). Science Mapping: A Systematic Review of the Literature. *Journal of Data and Information Science*, 2(2), 1-40.

He, J., Lou, W., & Li, K. (2019). How were science mapping tools applied? The Application of science mapping tools in LIS and non-LIS domains. *Proceedings of 82nd Annual Meeting of the Association for Information Science & Technology* (pp. 404-408). Maryland, USA: ASIS&T.

Comparing citations, usage, and interdisciplinarity in cover and non-cover papers

Ling Kong[1] and Dongbo Wang[2]

{[1] 2019214002, [2] db.wang}@njau.edu.cn
Nanjing Agricultural University, School of Information Management, Library Information and Archives
Management, WeiGang No.1 Nanjing, 210095 (China)

Introduction

Academic journal covers are important tools for attracting reader attention. Major journals regularly promote covers and featured papers, resulting in high traffic on their websites (Wang et al., 2014). Citation count is a prominent indicator of a paper's influence (Didegah & Thelwall, 2013), considered a standard, objective, and straightforward measure (Yan et al., 2012). Article usage data allow direct exploration of usage preferences (Wang et al., 2016; Chen, 2017). Usage reflects users' extensive attention behaviour, and summarises two types of use data: (1) times downloaded and (2) times citation exported. We obtained U1 (last 180 days) and U2 (2013–present) usage information.

After comprehensive consideration, we selected papers in Nature as our research objects. This study aims to address the following: What are the differences between cover and non-cover papers in terms of (1) citations, (2) usage, and (3) subject areas and research terms?

Data and Methods

To ensure a specific citation time window, the publication time interval was set to 2011–2015. The dataset collection work, completed in February 2021, yielded 503 cover papers and 5356 non-cover papers. A Python program crawled nature.com article pages for article subjects in physical sciences, earth and environmental sciences, biological sciences, health sciences, and scientific community and society (see Table 1 for number of subjects per discipline). Indicator data tags from WoS are publication year (PY), WoS Core Collection Citation frequency count (TC), U1, and U2. We analysed means, correlation coefficients, and interdisciplinarity. All statistical analyses were conducted using Python and Stata 15.1.

Table 1. Subjects in Nature

No.	Discipline	Disciplinary subjects
1	Physical sciences	576
2	Earth and environmental sciences	44
3	Biological sciences	1518
4	Health sciences	604
5	Scientific community and society	32
Total	5	2774

Results

Analysis of geometric means of citations and usage. The geometric means of TC, U1, and U2 were higher for cover papers than for non-cover papers (Figure 1); a yearly increase is observed in the difference between them. The TC and U2 of the 2012 cover papers are higher because of two important studies, a review of "A roadmap for graphene" (https://www.nature.com/articles/nature11458) and a paper on the healthy human microbiome (https://www.nature.com/articles/nature11234).

These results further show that Nature's papers conform to the citation rule: the best citation period is roughly 3–8 years.

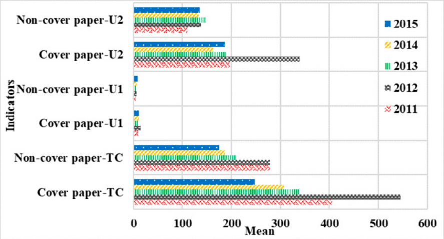


Figure 1. Annual geometric mean of TC, U1, and U2 of cover and non-cover papers.

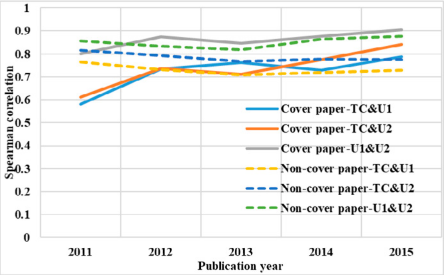
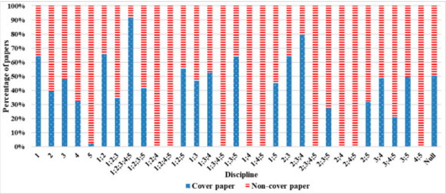


Figure 2. Annual Spearman correlations of TC, U1, and U2 of cover and non-cover papers.

Correlation analysis of citations and usage (Figure 2). Analysis of Spearman correlation annual evolution trends among TC, U1 and U2 shows that the correlation coefficient of cover papers fluctuated significantly and almost higher; for non-cover papers it was relatively stable and almost lower.

Interdisciplinary papers include research terms involving two or more disciplines. First, papers in a single discipline comprise 53.68% of cover papers and 53.70% of non-cover papers. Papers in two or more disciplines comprise 41.95% of cover papers and 40.36% of non-cover papers. Percentages of papers in two and three disciplines were higher in cover papers, but the differences were not significant. The proportion of physical sciences papers in cover papers is higher. The proportion of interdisciplinary physical sciences and earth and environmental sciences and of earth and environmental sciences and biological sciences in cover papers is higher. Moreover, most cover papers have higher cross-disciplinary ratios than do non-cover papers (Figure 3).



Note:1=Physical sciences; 2=Earth and environmental sciences; 3=Biological sciences; 4=Health sciences; 5=Scientific community and society.

Figure 3. Interdisciplinary comparison of cover and non-cover papers.

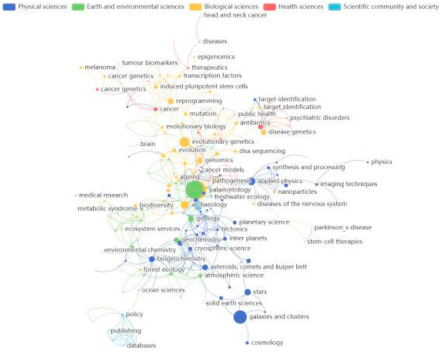


Figure 4. Interdisciplinary subjects of cover papers (Top 50, each discipline).

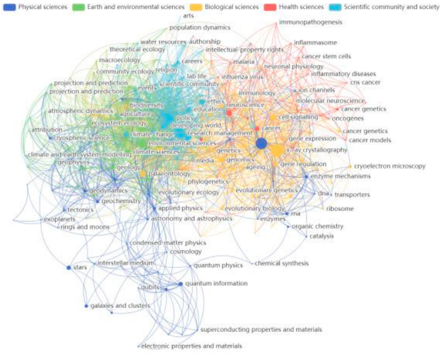


Figure 5. Interdisciplinary subjects of non-cover papers (Top 50, each discipline).

Second, there are significant differences in the Top 50 research subjects by discipline between cover and non-cover papers, and different subject clusters are apparent. Subject concentration and degree of subject crossover is higher in cover papers (Figure 4, 5).

Conclusions

Cover papers, as expected, gain more citations and usage compared to non-cover papers. Cover papers have unique high visibility and knowledge dissemination advantages. Within each discipline, certain research topics show interdisciplinary patterns. Inherently "better" articles are more likely selected as cover papers for their inherent quality and anticipated higher metrics and publicity (Phillips et al., 1991). In the future, we will analyse specific interdisciplinary research topics in cover and non-cover papers with different citation and usage rates. With frequent complexity, comprehensiveness, and integration problems, intersection and integration tendencies among different disciplines are increasing. However, it is necessary to analyse the specific knowledge crossover and fusion in the text content of crossover science.

Acknowledgments

The authors acknowledge the National Natural Science Foundation of China (Grant Numbers: 71774084) for financial support.

References

Chen, B. (2017). Usage pattern comparison of the same scholarly articles between Web of Science (WoS) and Springer. *Scientometrics*, 115, 519-537.

Didegah, F. & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics*, 7(4), 861-873.

Phillips, D., Kanter, E., Bednarczyk, B. & Tastad, P.L. (1991). Importance of the lay press in the transmission of medical knowledge to the scientific community. *The New England Journal of Medicine*, 325(16), 1180-1183.

Wang, X., Fang, Z. & Sun, X. (2016). Usage patterns of scholarly articles on Web of Science: A study on Web of Science usage count. *Scientometrics*, 109, 917-926.

Wang, X., Liu, C. & Mao, W. (2014). Does a paper being featured on the cover of a journal guarantee more attention and greater impact? *Scientometrics*, 102, 1815-1821.

Yan, R., Huang, C., Tang, J., Zhang, Y. & Li, X. (2012). To better stand on the shoulder of giants. Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries - JCDL 12, 51-60.

Knowledge Positioning of Patent Assignees

Chung-Huei Kuan [1] and Dar-Zen Chen [2]

[1] *maxkuan@mail.ntust.edu.tw*
Graduate Institute of Patent, National Taiwan University of Science and Technology, No. 43 Sec. 4 Keelung Rd., Taipei, Taiwan (R.O.C.)

[2] *dzchen@ntu.edu.tw*
Department of Mechanical Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan (R.O.C)
Center for Research in Econometric Theory and Applications, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan (R.O.C)

Introduction

We often read from articles in trade magazines and academic journals that an organization is referred to as a technology leader or follower. These terms are used by the article authors either quite arbitrarily or based on some proxy measures such as number of patents and their forward citation counts. This study is intrigued by the idea that whether this kind of *positioning* can be achieved *quantitatively* and more directly, rather than through proxies or speculation.

Related works may be categorized into two main approaches. The first approach positions organizational entities in a densely/strongly connected core or a loosely/weakly connected periphery following the core/periphery model proposed by Borgatti and Everett (2000). The other approach positions the entities relative to an evolutionary trajectory of the field (cf. Bekkers & Martinelli, 2012; Kim, Lee & Kwak, 2017), where the trajectory is unanimously derived using main path analysis (MPA) (Hummon & Doreian, 1989).

This study also follows the second approach for patent assignees but differs from the prior works in two main respects. Firstly, this study only requires that there are a number of representative patents reflecting knowledge evolution, regardless of how they are derived. Secondly, each assignee is associated with a number of position attributes, reflecting its multi-faceted positioning characteristics, thereby offering finer and more comprehensive analysis.

Position attributes

By considering that a technology field's patent citation network embodies a knowledge system for the field, and that a series of MS patents constitute the evolution trajectory, each patent of the technology field is identified to be at one of five possible positions relative to the series of representative patents: *mainstream* (MS), *forward and backward reachable* (FBR), *backward reachable only* (BR), *forward reachable only* (FR), and *non-reachable* (NR), based on whether they are on the trajectory and their reachability to and from the representative patents.

For the representative patents reflecting the field's knowledge evolution, they are said to be at MS positions, such as those denoted by the black nodes of Figure 1. A NR patent is one whose node cannot reach the nodes of the MS patents and vice versa, such as those denoted as white nodes in Figure 1. A FR patent is one whose node cannot reach those of the MS patent but may be reached by at least a node of the MS patents, such as those denoted by nodes 1-3 and 11-13. BR patents, denoted as nodes 16-17 and 25-27, are those whose nodes may reach at least one node of the MS patents but not the other way around. A patent whose node is both backward reachable from and forward reachable to at least one node of the MS patents, such as the node 15 is referred to as a FBR patent.

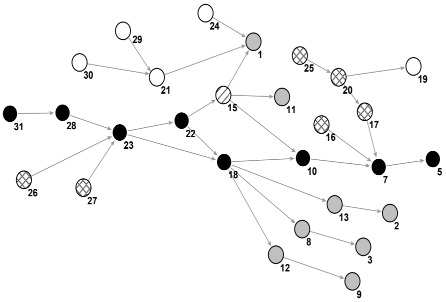


Figure 1. A portion of a real patent citation network.

After the positions of an assignee *i*'s patents are identified, assignee *i* can be characterized by five *position attributes* (MS_i, FBR_i, BR_i, FR_i, NR_i), which are shares of its patents at the respective MS, FBR, BR, FR, and NR positions.

Assignee classification

An assignee can be classified as a *uniquely positioned* or *multiply positioned* assignee, based on its position attributes. There are five types of uniquely positioned assignees referred to as *trendmakers*, *onlookers*, *leaders*, *followers*, and *boosters*, whose definitions are listed in Table 1. Trendmakers are assignees owning at least one MS patent. Onlookers are those having only NR patents, and their patents are completely irrelevant to the field's evolution. Leaders are those having only BR and NR patents, and these BR patents must lead ahead of and furnish knowledge to one or more MS patents. Follower*s* have only FR and NR patents, and their FR patents must lag behind and draw knowledge from at least one MS patent. Boosters have only FBR and NR patents, where their FBR patents not only draw knowledge from some earlier MS patents but also feedback knowledge to some later one.

Table 1. Position attributes for uniquely positioned assignees.

	MS	FBR	BR	FR	NR
Trendmakers	> 0	-	-	-	-
Boosters	0	> 0	0	0	-
Leaders	0	0	> 0	0	-
Followers	0	0	0	> 0	-
Onlookers	0	0	0	0	1

*'-'stands for 'don't care.'

There are four types of multiply positioned assignees as defined in Table 2. These assignees have mixed FBR, BR, and FR patents and, unlike the uniquely positioned, they assignees cannot be classified as having a specific position.

Table 2. Position attributes for multiply positioned assignees.

	MS	FBR	BR	FR	NR
Booster/ leader	0	> 0	> 0	0	-
Leader/ follower	0	0	> 0	> 0	-
Booster/ follower	0	> 0	0	> 0	-
Booster/ leader/follower	0	> 0	> 0	> 0	-

*'-'stands for 'don't care.'

Case study

27,161 U.S. biochip utility patents granted before 2019/3/31 are collected as case study, where the field's evolution trajectory is derived using MPA (with SPNP weight assignment and global search method). It is found that each of the 5,921 assignees involved (after disambiguation) can be classified effectively into exactly one of the uniquely and multiply positioned categories, as shown in Table 3.

Summary and Limitation

In this study, patent assignees are associated with positioning attributes through which they are classified into categories of distinct characteristics. A case study using MPA confirms the validity and simplicity of the proposed method. A major drawback of the proposed method, which will be dealt with in future study, is that it ignores the distance (e.g., number of "hops") to the MS patents. Therefore, the proposed method cannot tell whether some assignees are more directly influenced by or more contributing to the MS patents than others.

Table 3. Distribution of biochip patent assignees.

	Categories	Share
Uniquely positioned assignees 5,768 (97.42%)	Trendmakers	25 (0.42%)
	Boosters	2 (0.03%)
	Leaders	271(4.58%)
	Followers	1,124(18.98%)
	Onlookers	4,346 (73.40%)
Multiply positioned assignees 153 (2.58%)	Booster/ leader	3 (0.05%)
	Leader/ follower	118 (1.99%)
	Booster/ follower	12 (0.20%)
	Booster/ leader/follower	20 (0.34%)

Acknowledgments

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 110L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by the Ministry of Science and Technology (MOST), Taiwan, under Grant No. 110-2634-F-002-045-.

References

Bekkers, R. & Martinelli, A. (2012). Knowledge positions in high-tech markets: Trajectories, standards, strategies and true innovators. *Technological Forecasting and Social Change*, 79(7), 1192-1216.

Borgatti, S. P. & Everett, M. G. (2000). Models of core/periphery structures. *Social networks*, 21(4), 375-395.

Hummon, N. P. & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social networks*, 11(1), 39-63.

Kim, D. H., Lee, H. & Kwak, J. (2017). Standards as a driving force that influences emerging technological trajectories in the converging world of the Internet and things: An investigation of the M2M/IoT patent network. *Research Policy*, 46(7), 1234-1254.

The lockdown effect: Research in the age of a pandemic

Barbara S. Lancho Barrantes[1] and Nick Sheppard[2]

[1] *b.s.lancho-barrantes@leeds.ac.uk*
University of Leeds, Leeds, LS2 9JT (United Kingdom)

[2] *n.sheppard@leeds.ac.uk*
University of Leeds, Leeds, LS2 9JT (United Kingdom)

Introduction

The goal of this paper is to analyse how research has been undertaken during the Coronavirus pandemic. Scientific production about Covid-19 has increased rapidly along with conventional reviews, research synthesis studies and bibliometrics analyses. This paper is not about Covid-19 per se, rather scientific research in general during the pandemic. What types and volumes of research outputs have institutions, countries and disciplines produced during this time?

Research questions

• What research has been produced during 2020?
• What relationships are there between research entities e.g. number of publications/grants received?
• Which research institutions and countries have been the main actors in research during this period?
• Which disciplines have produced the most outputs?
• Has there been a greater number of open access publications than in the previous year?
• Have more preprints been shared in 2020?

Data and methods

The Dimensions database was used to collect data. We have focussed on the year 2020. Dimensions was used rather than Web of Science or Scopus as it contains the types of data we needed for this study: grants, datasets, publications (including preprints), citations, clinical trials and patents. The searches for retrieving the data were last updated on 5 Jan 2021.

Results

There was no significant increase in the number of publications produced in 2020 compared to previous years. In fact, there was higher growth in 2019 compared to 2018 than in 2020 compared to 2019. Although this could be affected by the fact that it has been possible to produce less in general but more COVID related which could make to balance the total amount of scientific production. A future analysis could analyze this circumstance in more detail to expose the reasons.

The Publications score shows high statistical correlation for Citations (0.940), Grants (0.90) and Clinical trials (0.82). Citations are highly correlated with Patents (0.83) and Grants (0.798). The Citations metric is also correlated with Datasets (0.774).

Table 1. Correlation matrix (Spearman) of content types.

Variables	PUBLICATIONS	CITATIONS	DATASETS	GRANTS	PATENTS	CLINICAL TRIALS	POLICY DOCUMENTS
PUBLICATIONS	**1**	0.940	0.749	0.909	0.811	0.826	0.335
CITATIONS	0.940	**1**	0.774	0.798	0.833	0.709	0.246
DATASETS	0.749	0.774	**1**	0.816	0.660	0.577	0.302
GRANTS	0.909	0.798	0.816	**1**	0.728	0.824	0.357
PATENTS	0.811	0.833	0.660	0.728	**1**	0.723	0.097
CLINICAL TRIALS	0.826	0.709	0.577	0.824	0.723	**1**	0.280
POLICY DOCUMENTS	0.335	0.246	0.302	0.357	0.097	0.280	**1**

Results indicate that the main types of research objects are highly correlated, indicating that the entire research ecosystem from inputs (grants) to outputs (publications) are interconnected.
N.B. correlation does not necessarily mean causation, there may be additional factors involved.

Comparing scientific production

The US and China have the highest number of publications, citations, datasets, grants, patents and clinical trials. Along with the UK they have also received the highest number of citations. Japan has more grants but does not generate more research outputs compared to other countries.
Harvard University, the University of the Chinese Academy of Sciences and the University of Toronto produced the most research. However Huazhong University of Science and Technology (HUST), Tsinghua University (THU) and the University of Oxford are have received the greatest academic impact. The University of Tokyo (UT), University of São Paulo and Shanghai Jiao Tong University have received the largest number of Grants.

Fields of research

The fields with the highest number of publications during 2020 were: Medical and Health Sciences (1,411,939), Engineering (721,491) and Biological Sciences (402,377). These have also received the highest number of Grants. However Biological Sciences (108,699), Information and Computing Sciences (100,270) and Environmental Sciences (57,752) are those with the highest number of Datasets.
Medical and Health Sciences, Engineering and Biological Sciences have both the highest number of publications and the highest number of Grants. There are fields e.g. History that have received a greater number of Grants but have published comparatively

fewer outputs. We take into account the different publication patterns across different disciplines.

Open access and preprints

Through the pandemic, researchers have embraced open publishing platforms and preprint servers to share their findings.

Table 2. Comparison Articles and Preprints

	2020- >2020	2019	2020- >2020	2019
	Articles		**Preprints**	
	4,501,885	4,274,495	401,375	271,830
Academic impact				
Citations	3,620,290	9,803,129	120,662	52,668
Citations (mean)	0.8	2.29	0.3	0.19
Publications with citations %	22.5	44.13	6.9	8.06
Altmetric Attention				
Publications with attention %	26.37	25.28	53.22	50.54
Altmetric Attention Score (mean)	15.97	12.74	11.04	6.77

The growth of preprints has been greater than articles, however articles have received higher citations. Total citation rate of preprints has increased presumably associated with time lag for citation in peer reviewed literature. This may indicate that preprints are citing other preprints. The Altmetric attention score is higher for articles than preprints but Preprints have more publications with Altmetric attention.

Table 3. Comparison by different type of publications

OPEN ACCESS	Publications	Citations	Citations mean	Publications with citations %	Publications with attention %	Altmetric Attention Score (mean)
Closed	3,111,542	1,513,034	0.49	18.64	19.04	9.54
All OA	2,856,997	2,350,072	0.82	18.33	29.56	18.92
Gold	2,283,109	1,900,341	0.83	18.56	24.21	21.75
Green, Accepted & Submitted	391,495	187,038	0.48	15.13	52.75	8.31
Green, Published	182,393	262,693	1.44	22.35	46.77	26.3

The number of closed publications exceeds that of open access. However, All OA has a higher citation mean (0.83) than Closed (0.49). Publications made open access via the green route have received the highest rate of citations per publication (1.44).

In Medical and Health Sciences, Biological Sciences and Physical Sciences more articles have been published open access than closed. Engineering, Information and Computing Sciences and Chemical Sciences have published more closed than open access. Citations per document from OA publications are higher in the Medical and Health Sciences (1.87) than from Closed publication (0.52), however Engineering has received more citations to Closed publications (1.03).

The following graph shows three variables, number of datasets, publications and citations. The US and China have the largest number of Datasets and Publications. However, despite fewer publications and datasets, the United Kingdom and Italy surpass them in citations per publication.

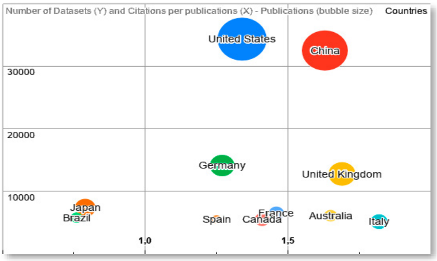


Figure 1. Comparison of datasets per countries

Conclusions

Our results suggest the following:

- Total scientific production in 2020 was 5,949,189 research outputs (Dimensions)
- The fields with the most publications are Medical and Health Sciences, Engineering, and Biomedical Sciences
- The three countries that have produced the most research are the US, China and the UK
- Research activity has not been unduly affected by Covid-19
- OA publications have higher citations (mean) than the closed publications. In particular via the Green route
- The growth of Preprints has been greater than for Articles.
- As expected the entire research system from inputs (grants) to outputs (publications) is interconnected. It is worth clarifying that not all research begins with a grant.

References

Haghani, M., Bliemer, M. C. J., Goerlandt, F. & Li, J. (2020). The scientifc literature on coronaviruses, covid-19 and its associated safety-related research dimensions: A scientometric analysis and scoping review. *Safety Science*, 129, 104806.

Barbisch, D., Koenig, K. & Shih, F. (2015). Is there a case for quarantine? Perspectives from SARS to Ebola. Dis. Med. Pub. *Health Prepar*. 9, 547-553.

Li, L. Z. & Wang, S. (2020). Prevalence and predictors of general psychiatric disorders and loneliness during COVID-19 in the United Kingdom. *Psych. Res*. 291, 0165-1781. doi: 10.1016/j.psychres.2020.113267

Recognition the Citation Author from JASIST Full Text

Zongyi Li[1], Dongbo Wang[2] and Si Shen[3]

[1] *zongyili@stu.njau.edu.cn,* [2] *db.wang@njau.edu.cn*
Nanjing Agricultural University (China)

[3] *shensi@njust.edu.cn*
Nanjing University of Science and Technology (China)

Introduction

Named-entity recognition is a fundamental and critical task in natural language processing, which is widely used in relation extraction, question and answering systems, semantic analysis, machine translation, etc. The current main approaches to named-entity recognition can be grouped into 3 categories: rule-based approaches, statistical-based approaches, and deep-learning-based approaches. Among the statistical-based approaches, CRF (Conditional Random Field) performs the best. Han et al. (2013) conducted a study on Chinese named entity recognition using the CRF model based on the characteristics of Chinese. In recent years, deep-learning-based approaches have become the mainstream approach for named- entity recognition research. Ali et al. (2019) used a method that combines multi-attention layer mechanism with LSTM (Long Short-Term Memory) for Arabic named-entity recognition. Jia and Xu (2018) proposed a CNN-BiLSTM-CRF based neural network model to simplify the Chinese named-entity recognition task. Li et al. (2020) used BERT (bidirectional encoder representations from transformers) and LSTM-CRF models to deal with the problem in Chinese clinical named entity recognition.

Citation author is the author whose work has been cited in an article. There are many studies on the subject of author co-occurrence network, however, recognition of citation author in academic texts is still insufficiently studied. Therefore, this study applies BERT and CRF models to recognizing citation author in academic texts, which may be helpful for subsequent analysis of author co-occurrence and deduction of author's similar research domains.

Data source and methods

Data source

In this study, data were from 696 published articles in *Journal of the Association for Information Science and Technology* (*JASIST*). These articles were published during the period 2016-2020.

After manual labelling, the citation author entities in JASIST full text are tagged. Some of the sentences that labelled with the citation author are shown below.

"An ontology is a semantic scheme that comprises the main classes (concepts) of a given domain of knowledge, their properties, interrelationships, and instances (<Citation Author> Noy </Citation Author> & <Citation Author> McGuinness </Citation Author>, 2001)."

"<Citation Author> Magatti </Citation Author>, <Citation Author> Calegari </Citation Author>, <Citation Author> Ciucci </Citation Author>, and <Citation Author> Stella </Citation Author> (2009) introduced an approach for labeling topics that relies on two manually labeled hierarchical knowledge resources: the Google Directory and the Open Office English Thesaurus."

The position of the 'citation author' appearing in the sentence can be roughly divided into two categories: one appears in parentheses at the end of each sentence, while the other appears more often at the beginning of the sentence.

Methods

The BIES (Begin, Inside, End, Single) method (Zhenget al., 2017) was used to label the citation author , while the other words were labelled by O. Specifically, citation author containing only one word are labeled as ywzz-S. For citation author consisting of two words, the first word is labeled with ywzz-B while the last word is labeled with ywzz-E. In addition, for citation author consisting of more than two words, they are labeled with ywzz-B, ywzz-I, and ywzz-E for the first, middle, and last words, respectively.

In this paper, the pre-training language model BERT is used, while the machine learning model CRF is used as a comparison experiment.

In the field of deep learning, BERT is a model with a bidirectional transformer structure proposed by Devlin et al. (2019), it applies attention mechanism to mine relationships between inputs and outputs. And In this work, citation author in JASIST texts are recognized by BERT pre-training model.

As for CRF, which belongs to the field of machine learning, is a probabilistic undirected graph discriminative model that solves the problem of labeling bias in sequence labeling for HMM and MEMM models. In this article we have used the open source toolkit CRF++0.58 and set NUM in -f

NUM to 2 (the default value is 1), as in tests -f 2 works better than the default value.

Results

A large number of pre-experiments were performed in the training of the BERT model. The final pre-training model with 12 layers, 748 hidden units, and 12 self-attention heads is applied in this paper. The values of the hyperparameters are as listed below: maximum sequence length 128, batch size 32, learning rate 2e-6, number of epochs 3.0, case insensitive. The reason for setting the hyperparameters in this way is that we found reducing the learning rate was effective in increasing the F1-value, while expanding the maximum sequence length had almost no effect on the F1-value. In addition, the effect when epochs = 10.0 is rather less effective than that when epochs = 3.0.

Then the 10-fold cross-validation tests were performed with BERT and CRF models respectively. The performance of each model was evaluated by P, R and F1 values, and results are as shown in Table 1 in summary.

Table 1. Results of Ten-Fold Cross Validation

NUM.	BERT	CRF
1	93.38%	91.62%
2	93.07%	90.32%
3	95.03%	90.96%
4	87.37%	85.03%
5	96.12%	92.88%
6	94.82%	93.25%
7	97.10%	94.38%
8	91.53%	81.47%
9	87.33%	62.99%
10	89.35%	61.68%
AVG_F1	92.51%	84.46%
MAX_F1	97.10%	94.38%

According to the above data, we plotted the maximum F1value and the average F1 value of the ten trials as histograms displayed in Figures 1.

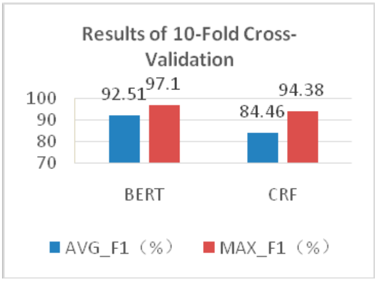


Figure 1. Histogram of average F1 value and maximum F1value of two models (BERT and CRF) in 10-fold cross-validation tests

It can be seen that the results of BERT are higher than those of CRF, regardless of the maximum F1value or the average F1 value. Overall, the average F1 values of both BERT and CRF are higher than 80%, indicating that the recognition effect of both is not generally satisfactory. Moreover, the optimal effect of BERT is better than that of CRF.

Conclusions

This study employed the BERT model and CRF model to extract the name entities of authors from the academic full text of JASIST (2016-2020). The average F1 values of both models is above 80%, and the optimal F1 value 97.10% is obtained by BERT. Although the recognition effect is acceptable, there is still room for improvement in this paper. In the future, to explore the distribution of citation author names in articles and better serve author co-occurrence studies, standardizing entity annotation, making the name of the same author consistent in expression and combining research questions of the citations, research methods of the citations are feasible solutions.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant Numbers: 71974094) .

References

Han, A. L. F., Wong, D. F. & Chao, L. S. (2013). Chinese Named Entity Recognition with Conditional Random Fields in the Light of Chinese Characteristics. *Springer Berlin Heidelberg*, 57-68.

Jia, Y. & Xu, X. (2018). Chinese Named Entity Recognition Based onCNN-BiLSTM-CRF. *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS),* 1-4

Ali, M. N. A., Tan, G. & Hussain, A. (2019). Boosting Arabic Named-Entity Recognition With Multi-Attention Layer. *IEEE Access*, 7, 46575-46582.

Li, X., Zhang, H. & Zhou, X. (2020). Chinese clinical named entity recognition with variant neural structures based on BERT methods, *Journal of Biomedical Informatics*, 107, 103422.

Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P. & Xu, B. (2017). Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme. *ArXiv, abs/1706.05075.*

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.

Recognizing Sentences Concerning Research Questions from the Full Text of *JASIST*

Chang Liu[1], Dongbo Wang[2] and Si Shen[3]

{[1]2020114020, [2] db.wang}@njau.edu.cn
Nanjing Agricultural University (China)

[3] shensi@njust.edu.cn
Nanjing University of Science and Technology (China)

Introduction

In the knowledge mining of academic full texts, the sentences about the research question usually indicate the research purpose and research topic of the paper, and have the function of leading the whole text and revealing the experimental content. Extracting this kind of content from literature is of great value for structural analysis, content analysis and bibliometrics analysis of texts. In terms of textual structure, these sentences tend to appear in the first part of academic texts. In terms of paragraphs, these sentences mainly appear in the abstract and introduction, but less frequently in other paragraphs. However, there are few researches on the recognition of research contents in academic texts. At present, manual and rule-based matching methods are mainly used to extract sentence entities in academic texts. For example, Chu (2017) analyzed the distribution of research methods in different academic journals by means of manual annotation. Gupta (2011) creates vocabulary matching rules for the entities in the paper so as to extract concerning sentences. However, the former of these two methods needs to spend a lot of manpower and time, while the latter needs to manually construct rules from a large number of annotated texts, which is faced with the problem of low accuracy and comprehensiveness of extraction. In this paper, machine learning and deep learning methods are adopted respectively, and three models, SVM, Bert and SciBERT, are adopted to recognize the sentences related to the research questions in *JASIST* academic papers. We compare the effect of three models in identifying sentences, and selects the model with the best effect to provide methods and ideas for subsequent research.

Dataset & Method

Dataset

The data used in this paper came from 502 full-text papers which published in *JASIST* Journal from 2017 to 2020.A total of 19 types of entities in the data set were annotated by manual annotation method, and all sentences labeled as research questions were extracted. The specific information of the corpus is shown in Table 1.

Table1. Basic Information of our Corpus

No.	Item	Count
1	Papers	502
2	Total Sentences	54,479
3	research questions sentences	1,940
4	Marked sentences in each article average number	3.86
5	Average words number in each sentences	27.99
6	The longest sentence words number	255

Sentences labeled as research questions usually contain a description of the specific research object, a description of the specific purpose of the research, and a plan of the experimental process. The contents of some annotated sentences are shown in Table 2.

Table2. The Information of some annotated sentences

Reasonsfor annotating	Annotated example
Research purpose	…,The primary goal of this study was to start to ascertain what impacts the development of … .
Description of the subject	We study the problem of how to detect the temporal patterns of …. .
Planning of theexperimental process	RQ2： What are the spatial（physical） characteristics of indoor web use？

Methods

Our experiment belongs to the text classification task in natural language processing. In this experiment, SVM (Support Vector Machine) was used as machine learning model, and Bert and SciBERT were used as deep learning models to verify the recognition effect of traditional machine learning methods and deep learning methods on data sets. SVM model is dedicated to finding the optimal hyperplane, which can separate the training samples and make the differences between different classes as obvious as possible, and in this experiment, TF-IDF is used to vectorize text. Bert (Devlin, 2018) is a pre-training model open source by Google in 2018. This model is based on multi-layer bidirectional transformation and decoding, and relies on two processes, pre-training and fine-tuning, to perform natural language processing tasks. The model achieved the best performance in 11 different NLP tests, which makes it one of the most effective models in NLP. SciBERT (Beltagy, 2019) is a pre-training model based on the Bert framework and trained by the academic paper corpus containing 1.14 million academic papers. Scibert built a new vocabulary based on the scientific corpuses, which overlapped 42% with the vocabulary of Bert-base. The effect of this model is better than Bert indownstream tasks of identifying academic text. However, this model has not been tested on corpus related to LIS.

Result

Based on a large number of preliminary experiments, the pre-training model of 12 layers, 768 hidden units,12self-attention heads and 110M parameters published by Google was selected. In the experiment, the same parameters were set for both Bert and SciBERT: epoch is 10, max-seq-length is 512, batch size is 32, and the learning rate is 2e-5. The parameters of the SVM model used are set as C= 2.0,kernel= 'RBF', gamma= 0.5. Our data set is divided into a training set and a test set in a 9:1 ratio. Three models were respectively used for 10-fold cross-validation of experimental data. P (accuracy rate), R (recall rate) and F (harmonic mean value) were used as evaluation indexes of the model, and the final results were shown in Table 3.

Table 3. Results of 10-Fold Cross-Validation on each model

Model	MAX-F or AVG-F	Non-Research QuestionsF-value	Research Questions F-value	Micro-AVG F-value	Macro-AVG F-value
SVM	MAX-F	98.09%	25.40%	95.29%	61.74%
	AVG-F	97.33%	18.18%	93.36%	57.75%
BERT	MAX-F	97.60%	56.35%	95.45%	76.97%
	AVG-F	97.07%	42.54%	94.44%	70.45%
SCiBERT	MAX-F	97.78%	59.54%	95.80%	78.86%
	AVG-F	97.78%	53.64%	95.76%	75.81%

The result shows that the SVM model recognition effect is poor. The recognition effect of Bert is not good, and the recognition effect of SciBERT is the best. SciBERT outperformed traditional machine learning models by nearly 20 percentage points. In the classification result, the maximum F value is 78.86%. The optimal result in the experiment is obtained by using SciBERT, which indicates that the use of pre-training method can effectively improve the classification effect on the data set of a specific domain.

Conclusion

In this paper, machine learning and deep learning methods are used to study the sentences of research questions in academic full-text. From the results, the effect of deep learning model is better than statistical learning model, and model pretrained with specific corpus has better performance than the original model in the downstream tasks of the same domain. This study provides a way to identify the content related to the research questions in academic texts. This method can quickly locate and extract the sentences related to the research topic, questions and purpose, so as to quickly understand the knowledge ontology, clarify the research hotspot and construct the knowledge graph. This study provides a new perspective for the bibliometrics analysis based on the full text, and how to improve the recognition effect of the model will be the focus of the next research.

Acknowledgments

The authors acknowledge the National Natural Science Foundation of China (Grant Numbers: 71974094) for financial support.

References

Chu, H. T. & Ke, Q. (2017). Research methods: What’s in the name?. *Library & Information Science Research*, 369(4), 284-294.

Gupta, S. & Manning, C. (2011). Analyzing the dynamics of research by extracting key aspects of scientific papers. *Proceedings of the5th International Joint Conference on Natural Language Processing. Asian Federation of Natural Language Processing*

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding *.arXiv preprint arXiv: 1810. 04805v1*

Beltagy, I., Cohan, A. & Lo, K. (2019). SciBERT: Pretrained Contextualized Embeddings for Scientific Text. *ArXiv, abs/1903.10676.*

Research Methods Sentence Identification in Academic Full Text Based on Deep Learning

Jiangfeng Liu[1], Dongbo Wang[1,2] and Si Shen[3]

{liujf, db.wang}@njau.edu.cn, shensi@njust.edu.cn
[1]Nanjing Agricultural University (China)
[2]KU Leuven (Belgium)
[3]Nanjing University of Science and Technology (China)

Introduction

Informetrics is a science for studying the value of literature. Informetrics in its broad sense includes bibliometrics, scientometrics, informetrics, webmetrics and altmetrics. However, the development of informetrics has not been free from the framework of using external indicators to measure research value. With the further development of computer technology, the natural language processing technology based on machine learning and deep learning provides new ideas for the development of informatics from the aspect of information internal characteristics.

In recent years, the study of academic full-text has attracted wide attention. Extracting specific types of entities and sentences from texts, such as software entities, algorithm entities, time entities, and research prospect sentences, has become the object of relevant research.

The research method in this research refers to the methods, tools, means, techniques and schemes for solving the problems in the application domain. The section of dataset and method of an academic article usually contains a series of sentences that introduce the research methods it used. These sentences simply and directly indicate the specific methods of the research, and are of great significance for understanding the ideas of the research, repeating previous studies and understanding the progress of the discipline. Research method can measure the level and value of the research to some extent. Previous studies have been carried out on the extraction of specific sentences. Wang and Zhang (2020) build a dictionary of algorithms by manually annotating the contents of papers, and sentences containing algorithms in the dictionary are extracted through dictionary-based matching. However, the research method has greater uncertainty than the algorithm, so the method of matching by building dictionaries does not work. This paper uses machine learning and deep learning models to identify the sentences of the research method.

Data & Method

The data used in this research was from the Journal of Association for Information Science and Technology (JASIST). A total of 502 articles containing 138197 sentences (including 4030 research methods sentence) published between 2017 to 2020 were downloaded and the full text of these articles were used as the corpus of this experiment.

The full text of articles was divided into sentences. We used label B to mark the sentences that belong to the research method and the label O was used for tagging sentences that did not belong to the research method. Sentences belonging to the research method have a lower proportion in all sentences. The following two methods were adopted to increase the proportion of sentences belonging to the research method in the experimental set. The first method is under-sampling. We eliminated sections that do not contain research methods. The second method is Data Augmentation. It is performed by translating sentences to other languages and back to English.

Model

In this study, the following two classification models were used to identify whether a sentence was a research method sentence. First, SVM model; second, BERT model. SVM (Support Vector Machine) is a classical generalized linear classifier that classifies data according to supervised learning. Its decision boundary is to find a hyperplane with maximum margin for the learning sample.

Bert is a language representation model introduced in 2018, which stands for Bidirectional Encoder Representations from Transformers (Devlin et al., 2018). Pre-training is already performed with Bert on linguistic representation before the training begins. The pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, without substantial task-specific architecture modifications. This experiment is based on the following two Bert pre-training models, one is (cased_L-12_H-768_A-12), and the second is the Scibert (scibert-scivocab-uncased), proposed by Beltagy (2018).

Experiment

In an experiment with the original corpus, the recognition effect of the classification model is not ideal. Therefore, in the 2nd stage experiment, we first take the section structure as a unit and remove the section which does not include the research method sentence. At this point, the research method sentences accounted for about 8% of the total sentences, which was still low. Therefore, we used Baidu translation tools to translate the sentences of the research results into Chinese, Cantonese, Korean, Japanese, etc., and then translated the results back to English. The final sentence is semantically the same or similar to the original sentence, so it can be considered as a research method sentence. The enhanced research method sentences accounted for about 10% of the total sentences. When training the corpus before and after the test data enhancement separately, the F-value increased by about 30%.

We divided the corpus according to the training set: validation set 9:1 based on 10-fold cross validation and evaluated the effect using F-value, Micro-AVG F-value and Macro-AVG F-value. In the SVM classification experiment, TFIDF was used to vectorize the text. The Summarized results is shown in Table 1.

Table 1. Summarized Results

Model or Pre-Model	Indicator	Non-Research MethodSentence F-value	Research Method Sentence F-value	Macro-AVG F-value	Micro-AVG F-value
SVM	Avg	95.37%	34.52%	89.24%	64.95%
	Max	95.72%	37.59%	90.06%	66.47%
	Min	94.87%	29.87%	88.14%	62.66%
Bert-Base	Avg	94.52%	34.82%	89.80%	65.95%
	Max	94.80%	39.19%	90.37%	68.27%
	Min	94.10%	29.18%	89.10%	63.00%
SCI-Bert	Avg	94.14%	35.39%	89.67%	65.58%
	Max	94.77%	37.68%	90.35%	67.35%
	Min	93.57%	34.32%	89.39%	64.92%

In general, the recognition effect of the Bert model especially the Scibertis better than that of the SVM model. After data enhancement, the recognition effect is greatly improved compared with that before data enhancement. With further random negative sampling, the effect of SVM model on the recognition of the research method sentence can reach 60%. We also spent a lot of time checking the corpus and the effectiveness of the model has also been improved. The experimental result shows that the number of research method sentences in the corpus of this experiment is still small, and we plan to continue to expand the corpus size, expand the proportion of research method sentences appropriately and reduce the cost of manual annotation based on active learning strategy.

Conclusions & Future Research

This study constructed an academic full-text corpus based on JASIST. By using SVM and Bert respectively, and using under sampling and data enhancement methods, the maximum harmonic mean value reaches 68.27% in the case of extremely uneven samples. This study provides an idea for identifying research method sentences in academic full-text research. Further, cluster analysis and keyword topic recognition can be carried out on the research method sentences according to the experimental results, so as to understand the internal characteristics of the research method sentences at a more microscopic level.

In addition, the research question sentence, the research result sentence, the research prospect sentence and the research method sentence are all important components of the academic full-text research, so it can be considered to construct a corpus containing the pre-annotated data of the four types of sentences, and at the same time carry out the classification experiment on them. Further research can focus on the mutual relations and internal characteristics of these types of sentences.

Previous study (Small, 2018) has classified articles into methods and their types and non-methods. This paper studied the relationship between citations and whether they are methodological papers which provided a further idea for our research. Next, we can categorize the identified sentences (research questions, methods, results, etc.) to summarize the underlying themes of the overview literature and compare and analyze them with their informometric indicators.

Acknowledgments

Researches are supported by the National Natural Science Foundation of China (Grant Number: 71974094).

References

Wang, Y. & Zhang, C. (2020). Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. *Journal of Informetrics*, 14, 101091-101091.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*NAACL-HLT*.

Beltagy, I., Cohan, A. & Lo, K. (2019). SciBERT: Pretrained Contextualized Embeddings for Scientific Text.*ArXiv, abs/1903.10676*.

Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *J. Informetrics*, 12, 461-480.

Does Technology of Climate Change Mitigation Transfer Follow UNFCCC?

Paoling Liu[1] and Chun-Chieh Wang[2]

[1] *r09630009@ntu.edu.tw*
Dept. of Bio-Industry Communication and Development, National Taiwan University, Taiwan (R.O.C.)

[2] *wangcc@ntu.edu.tw*
Dept. of Bio-Industry Communication and Development, National Taiwan University, Taiwan (R.O.C.)
Center for Research in Econometric Theory and Applications, National Taiwan University, Taiwan (R.O.C.)

Background

For the revolution of clean energy and environmental justice, the current US President Joe Biden vowed to return to the Paris Climate Agreement as soon as he took office, and emphasized that climate change is one of the main axes of US administration in the next few years (Biden, 2019). Since the United Nations Framework Convention on Climate Change (UNFCCC) forced the commitment of technology transfer in 1992, how to fulfill the commitment and effectively promote the international transfer of climate-friendly technologies in reality has been the theme of debate (Zhou, 2019). Since the Bali Roadmap (2007) further emphasized the importance of clean energy technology transfer (UNFCCC, 2008), the "Cancun Declaration" held in Mexico in 2010 placed more emphasis on climate change mitigation and adaptation, and responded to the technology transfer requirements of developing countries. The Paris Agreement in 2015 once again emphasized the commitment of all parties to strengthen climate technology cooperation and the requirements for clean energy technologies (Koskina, Farah, & Ibrahim, 2020; UNFCCC, 2016). In previous studies, it has been pointed out that patents can be used to analyze technology transfer in mitigating climate change (Goldar et al., 2019). However, this study did not explore whether countries have followed the technology transfer requirements mentioned in the Framework Convention on Climate Change.

The objective of this study was: Using patent data to explore whether there is knowledge spillover from developed countries to least developed countries. The completion of this study would help to understand if the mitigation mechanism mentioned in the Climate Change Convention remained consistent with the future outlook of the United Nations Framework Convention on Climate Change, and would also clarify any knowledge spillover from developed countries to least developed countries.

Methodology

This study adopted the citation analysis of patentometrics to measure knowledge spillover among countries. In order to measure the spillover effect of climate change mitigation technologies in various countries, the United States Patent and Trademark Office (USPTO) patent data was used to analyze the status of technology transfer between 129 countries, the code Y02 of International Cooperation Patent Classification was used for calculation. And in accordance with the United Nations Framework Convention on Climate Change to organize important meetings on mitigating climate change, the Cancun Convention held in 2010 and the Paris Agreement held in 2015, two periods of time were respectively classified at the following intervals: 2009 to 2014 and 2015 to 2020. A total of 263,912 patents were analyzed in the two time periods. The Two Mode Analysis in terms of social network analysis was used to identify the importance of each node in the network, and quantify the spillover effect of climate change mitigation technology in different countries, in order to understand the flow of technology spillover to see if there were any from developed countries to least developed countries.

Results

This study was an attempt to explain the spillover network of climate change mitigation technologies. According to the statistics of the World Bank (2020), the per capita GNI calculated by the Atlas method of the World Bank can be categorized as entities of low-income economies, middle-income economies, upper-middle economies and high-income economies. Figures 1 and 2 list the analysis of spillover network in two periods, in which the red circle represents high-income countries, the blue circle represents middle-and high-income countries, and the yellow circle is low-and middle-income countries. The green box node represents how the climate change mitigation technologies of various countries flow in different levels of economies, and the line arrows represent their knowledge flowing out to countries at different levels.

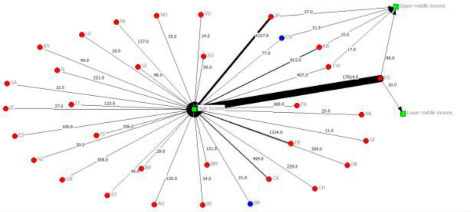


Figure 1. Clean Energy Technology Spillover Map (2009-2014)

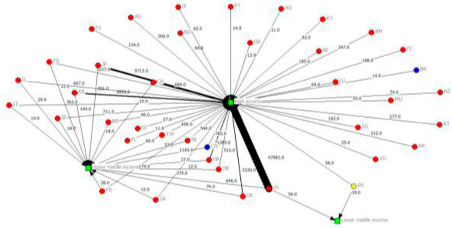


Figure 2. Clean Energy Technology Spillover Map (2015-2020)

According to Figure 1 from 2009 to 2014, climate change mitigation technologies remained mostly within exchange between high-income countries. According to Figure 2, technologies for mitigating climate change were still mainly exchanged within high-income countries, and to a lesser extent from high-income countries to middle-and high-income countries. From Figures 1 and 2, technologies to mitigate climate change mainly flowed in high-income countries and middle-and high-income countries. Compared with Figure 1, Figure 2 clearly shows, as in India, the technologies have begun to flow into low-and middle-income countries, except the United States, but such flow still remained quite low, not to mention the flow to low-income countries.

Conclusion

Using the results of this study, it was determined that climate change mitigation technologies did spill over in time, and the final trend of spillover of related patents and knowledge from countries could help predict the future development of this technology field. However, as found in this study, the knowledge and technology flow to mitigate climate change was still limited to high-income countries and middle-and high-income countries, and the knowledge spillover to low-and middle-income countries and low-income countries was minimal. Therefore, it remained an important topic to repeatedly emphasize the import implication of technology transfer in past and future conferences, that it is imperative to strengthen and actively promote technology development and transfer (Koskina et al., 2020; UNFCCC, 2016). The measures implemented have failed to achieve the necessary commitments mentioned in the Climate Change Convention; thus, the actions and outcomes were inconsistent with the future prospects of the UNFCCC. In term of research limitation, because this study only analyzed data from the U.S. Patent and Trademark Office, if the related technology has not been patented in the United States, it cannot be analyzed.

Acknowledgments

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 109L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by the Ministry of Science and Technology (MOST), Taiwan, under Grant No. 109-2634-F-002-045-.

References

Biden, J. (2019, January 15). *The Biden Plan For A Clean Energy Revolution And Environmental Justice.* https://joebiden.com/climate-plan/

Goldar, A., Sharma, S., Sawant, V. & Jain, S. (2019, July). *Climate Change & Technology Transfer–Barriers, Technologies and Mechanisms.* Indian Council for Research on International Economic Relations. http://hdl.handle.net/11540/10948

Koskina, A., Farah, P. D. & Ibrahim, I. A. (2020). Trade in clean energy technologies: sliding from protection to protectionism through obligations for technology transfer in climate change law, or Vice Versa? *The Journal of World Energy Law & Business*, 13(2), 114-128. https://doi.org/10.1093/jwelb/jwaa013

UNFCCC. (2008, March 14). *Report of the Conference of the Parties on its thirteenth session, held in Bali from 3 to 15 December 2007.* https://unfccc.int/resource/docs/2007/cop13/eng/06a01.pdf

UNFCCC (2016, January 29). *Report of the Conference of the Parties on its twenty-first session, held in Paris from 30 November to 13 December 2015.* https://unfccc.int/resource/docs/2015/cop21/eng/10a01.pdf

World Bank (2020, July 1). *World Bank list of economies.* https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

Zhou, C. (2019). Can intellectual property rights within climate technology transfer work for the UNFCCC and the Paris Agreement? *International Environmental Agreements: Politics, Law and Economics*, 19(1), 107-122. DOI: 10.1007/s10784-018-09427-2

Countries multidimensional scientometric performance profiles compared with a self-organized neural network

Ibis A. Lozano-Díaz[1], Ricardo Arencibia-Jorge[2] and Humberto Carrillo-Calvet[3]

[1] *ibis.alozano@gmail.com*
Faculty of Sciences, National Autonomous University of Mexico, Circuito Centro Cultural s/n, Coyoacan 04510, Mexico City (Mexico)

[2] *ricardo.arencibia@c3.unam.mx*
Complexity Sciences Center (C3), National Autonomous University of Mexico, Circuito Centro Cultural s/n, Coyoacan 04510, Mexico City (Mexico)

[3] *humbertocarrillo@ciencias.unam.mx*
Faculty of Sciences and Complexity Sciences Center (C3), National Autonomous University of Mexico, Circuito Centro Cultural s/n, Coyoacan 04510, Mexico City (Mexico)

Introduction

Country profiles reveal interesting aspects of their economic-demographic landscape and scientific research performance. In earlier studies, different authors have used the multidimensional approach to describe national research profiles (Glänzel, 2000; Glänzel, Leta, & Thijs, 2006; Leta, Glänzel, & Thijs, 2006). However, the comparison of national performance profiles has received less attention, perhaps due to composite indices widely used by prevalent rankings.

The current study avoids the rankings approach. The aim is to compare Mexico's scientometric performance with other countries' performance with similar economic demographic profiles. Since multidimensional profiles comparisons are involved, we resort to an artificial intelligence method based on a self-organized neural network (SOM) (Villaseñor; Carrillo-Calvet, & Arencibia-Jorge, 2017).

Materials and Method

We first identify countries with similar economic-demographic profiles to Mexico, and then we compare their scientometric performance. A sample of 108 countries was selected in the Web of Science to calculate four economic-demographic indicators: Population in 2019 (Pop 2019, World Bank Data), Global Competitiveness Index (GCI, World Economic Forum), Economic Complexity Index (Hidalgo & Hausmann, 2009) (ECI), and Gross domestic product per capita (GDP per capita 2019, World Bank Data). Comparing the multidimensional economic-demographic profiles of the 108 countries was carried out automatically using the SOM neural network implemented in the software system ViBlioSOM 2.0 (Jimenez Andrade, Villaseñor-García, & Carrillo-Calvet, 2019). Thus, the neural network identified in one cluster a set of 17 countries with the most similar profiles to Mexico.

To compare Mexico's scientometric performance with the performance of these other 17 countries, we used the Documents per 10000 inhabitants indicator (calculated taking into account data from Web of Sciences (InCites) and the World Bank) and other five indicators from InCites: % of international collaboration; % of Documents in Top 1% and Top 10% of most cited articles in their subject categories; % of Documents in Q1 Journals, and Category Normalized Citation Impact (CNCI). For the comparison of the scientometric profiles we used again the neural network to produce a clusters map that can be interpreted with the help of a collection of associated components maps. To calculate the optimal number of clusters, we used the Davies-Bouldin index (Davies & Boulding, 1979).

Results and Discussion

The neural network grouped the sample of 108 countries in seven clusters sharing the most similar economic-demographic characteristics (Figure 1).

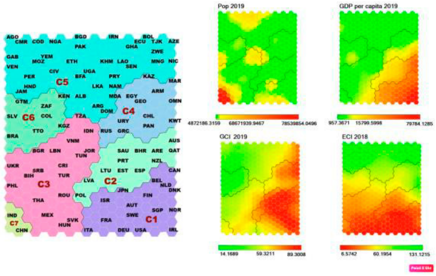


Figure 1. Comparison of 108 countries using economic-demographic indicators. Country codes use the ISO 3166-1 alpha-3 international standard.

Four component maps corresponding to each indicator are also displayed in the figure to interpret the clusters map. The countries with the most complex economies and the highest values of GDP per capita are grouped in clusters C1, C2 and C4.

In cluster C7 China and India appear as outliers due to their extremely high population. Eighteen countries belong to cluster C3, sharing high values of the Economic Complexity Index and medium values in the other indicators. This cluster includes two Latin American countries (Costa Rica and Mexico) and eight European countries (Bosnia and Herzegovina, Bulgaria, Turkey, Romania, Serbia, Hungary, Ukraine, Slovakia), four Asian countries (Indonesia, Philippines, Thailand and Viet Nam), two African (Tunisia and Tanzania), and two Middle East countries (Jordan and Lebanon). Considering that these 18 countries are the most similar to Mexico in terms of demographic indicators, we then compare their Scientometric multidimensional performance profiles using the neural network (Figure 2).

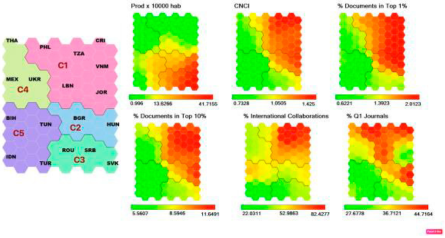


Figure 2. Multidimensional scientometric profiles of countries with similar economic-demographic profiles to Mexico (2017-2019).

Of the five clusters identified by the neural net, the countries with the best performance profiles are located in clusters C1 and C2. Countries in Cluster C2 and C3 stand out as the most productive ones and C1 have the highest international collaboration score. Mexico, Ukraine and Thailand belong to C4. This cluster together with C3 showed the lowest values in the productivity and impact indicators. However, C4 has a better percentage of articles in Q1. Meanwhile C3 has a better productivity level.

Focusing on Mexico`s relative scientometric performance as an example, it is clear from the maps in figure 2 that Mexico is outperformed in productivity and impact by the 11 countries which belong to clusters C1, C2 and C3. For instance, it has a score of 0.94% in documents in top 1% indicator vs a maximum of 2% stained by some countries in cluster C1 and C2; it has 6.69 in documents in top 10% indicator vs a maximum of 11.6 for some countries in these clusters. Similarly, in the Category Normalized Impact Factor, Mexico has 0.84; meanwhile countries with the best score reach 1.4. Regarding the productivity indicator, Mexico has five papers per 10000 habitants; meanwhile the larger score in cluster C3 is 42 and Tanzania has a minimum of 0.8.

Conclusion

The neurocomputational approach facilitated identifying 18 countries which makes sense to compare Mexico, taking into account demography and economic patterns, and comparing Mexican scientometric performance with these countries. Compared with countries with the same economic-demographic landscapes, the productivity and impact of Mexican output were lower than expected. Therefore, beyond the weight of demographic and economic variables, other factors must be analyzed in further studies, such as developing national academic journal ecosystems, the national research evaluation policies, or the level of democracy. Finally, we remark that to carry out the multidimensional comparisons needed by this investigation is not a trivial achievement. The visual output of the neural net provided a clear results' representation.

Acknowledgments

This research was supported by UNAM. We appreciate the collaboration of José Luis Jiménez Andrade and Javier García García (C3).

References

Davies, D. L. & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224-227

Glänzel, W. (2000). Science in Scandinavia: A bibliometric approach. *Scientometrics*, 48(2), 121-150.

Glänzel, W., Leta, J. & Thijs, B. (2006). Science in Brazil. Part 1: A macro-level comparative study. *Scientometrics*, 67(1), 67-86.

Hidalgo, C. A. & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26), 10570-10575.

Jimenez Andrade, J. L., Villaseñor-García, E. A. & Carrillo-Calvet, H. (2019). LabSOM: Self Organizing Maps Laboratory. https://doi.org/https://doi.org/10.5281/zenodo.3630581

Leta, J., Glänzel, W. & Thijs, B. (2006). Science in Brazil. Part 2: Sectoral and institutional research profiles. *Scientometrics*, 67(1), 87-105.

Villaseñor, E. A., Carrillo-Calvet, H. & Arencibia-Jorge, R. (2017). Multiparametric characterization of scientometric performance profiles assisted by neural networks: a study of Mexican higher education institutions. *Scientometrics*, 109(2), 77-104.

Analysis of Topic Evolvements of COVID-19 Scientific Literatures based on Probabilistic Modelling

Ma Lili[1], Zhao Yanqiang[1], Yue Mingliang[1,2,3] and Ma Tingcan[1,2,3]

{mall, zhaoyq, yueml, matc}@whlib.ac.cn
[1]Wuhan Library of Chinese Academy of Science, Wuhan (China)

[2]Department of Library, Information and Archives Management, University of Chinese Academy of Sciences, Beijing (China)

[3]Hubei Key Laboratory of Big Data in Science and Technology, Wuhan (China)

Introduction

Since it was first reported at the end of December 2019, the new coronavirus disease (COVID-19) is undoubtedly the most important health issue of global concern. With the increasing of research papers related to COVID-19, bibliometric analysis becomes a promising way for comprehensively understanding the global research status in the field (Al-Gheethi, 2020). In this paper, LDA model is adopt to recognize the topics of the bibliometric data of the COVID-19 related research paper. We proposed a weighted summation method to get full use of the probabilities reported by the model to analyze the monthly topic evolvements and discuss the potential reasons of the evolvements.

Data and Method

The bibliographic data of COVID-19 were searched and downloaded using TS = (2019-nCoV or SARS-CoV-2 or COVID-19 or COVID-2019 or "Coronavirus disease 2019") as search strategies from Web of Science Core Collection, Science Citation Index Expanded (SCI-EXPANDED) database . The article type was set to include article, letter and review. The data relates to 36913 research papers, among which 30122 papers have been officially published that can be mapped to the month of publication. In this paper, all the papers' bibliographic data is used in the topic modelling to get a more comprehensive topic recognition, while only the officially published papers is used in the evolvement analysis. Considering the publication delay, this paper mainly analyzes the topic evolvement from January to October, 2020.

In this paper, LDA model is used to recognize the research topics included in the COVID-19 related research papers. The model can be formally represented as Ω=Φ×Θ, where Ω, Φ and Θ is document-word distribution, topic-word distribution and document-topic distribution respectively, × represents matrix multiplication (Blei et al., 2003). Given the a set of *m* documents and a topic number *n*, the model returns *m* vectors of length *n* indicating to what extent each of the *m* papers are related to each of the *n* topics and *n* vectors of keywords that can be used to interpret the research topics.

Let $\Theta=\{\theta_{ij} \mid i\in[1,m], j\in[1,n]\}$, each θ_{ij} be the probability of paper *i* on topic *j* returned by LDA modelling. As afore-mentioned, the probability can be interpreted as the weight indicating to what extent a research is related to a topic. For each topic, we can add up the weights relating to the topic of all the papers to get an overall recognition of the focus on the topic in the current research. More formally, we can define $S_j=\sum_i \theta_{ij}$ as the Research Strength of topic *j* in the current research, and define $S_j^k=\sum_i \theta_{ij}, P_i=k$ as the research strength of topic *j* in the research of the *k*th month of 2020, where P_i represents the publication month of paper *i*. In the next section, we analyze the evolvement (of the research strength) of each COVID-19 related research topic.

In this paper, title and abstract of the papers' bibliographic data were extracted as the input of LDA model. Topic number *n* was set as 10 to 50 and the corresponding topic coherences were calculated to find the best topic number, where topic coherence is a quantitative measure that can be used to evaluate the quality of the topic modelling, with higher coherence indicating higher quality (O'Callaghan et al., 2015). The distribution of coherences of our case shows a maximum coherence at *n* = 25.

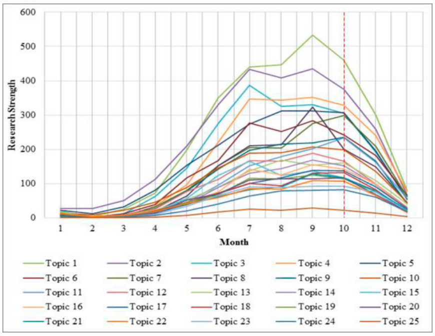


Figure 1. Topic evolvements of the 25 research topics

Table 1. The 25 topics

Topic
Pandemic and its impacts and challenges on medical, social, economic, etc.(Topic 1)
Infection and coinfection of coronavirus disease 2019 (Topic 2)
Guidelines in healthcare facilities and protection for healthcare workers (Topic 3)
Influence of basic disease on COVID-19 infection and mortality (Topic 4)
COVID-19 transmission model and epidemic development trend (Topic 5)
CT and other imaging diagnosis (Topic 6)
Study on psychological and mental problems during the COVID-19 pandemic (Topic 7)
The impact of the epidemic on surgery and education (Topic 8)
Treatment of critically ill COVID -19 patients (Topic 9)
Report and epidemiological characteristics of epidemic situation in worldwide (Topic 10)
The clinical characteristics and prediction indexes of patients with COVID-19 (Topic 11)
Drugs treatment and effects study (Topic 12)
Rapid diagnosis based on nucleic acid detection (Topic 13)
Immunological study for COVID-19 (Topic 14)
Cancer treatment during the COVID-19 pandemic (Topic 15)
Analysis of COVID-19 incidence and mortality among different regions (Topic 16)
Deep learning and AI in the research of COVID-19 (Topic 17)
SARS-COV-2 receptor and infection mechanism (Topic 18)
Systematic review and meta analysis (Topic 19)
The air and contact transmission of SARS-COV-2 and the role of protective equipment such as masks (Topic 20)
COVID-19 inhibitors target screening, molecular simulation and potential therapeutic drugs research (Topic 21)
COVID-19 genome sequence research and vaccine development (Topic 22)
SARS-COV-2 infection in pregnancy and maternal and neonatal vertical transmission (Topic 23)
The effect of the COVID-19 on air pollution and air quality (Topic 24)
Others (Topic 25)

Topic evolvement

Fig. 1 gives the time evolution of the research strength of each topic, and the 25 research topics are shown in Table 1. It can be seen that, from January to February, in the early stage of the epidemic, scientists around the world did not know much about this sudden novel coronavirus and its pathogenic mechanism. The timely sharing and disclosure of epidemic data, the prediction of epidemic development trend as well as the rapid detection and identification of COVID-19 patients played important roles in controlling the spread of the epidemic. Therefore, topic 1, 2, 5, 10, 13 that related to the infection and coinfection, challenge, rapid diagnosis of COVID-19, as well as the epidemiological characteristics and development of epidemic were the most concerned.

From March to July, the COVID-19 pandemic caused the significant influence worldwide, and how to protect healthcare workers became the focus, therefore, in addition to topic 1 and 2, prevention and control guidelines in healthcare facilities and protection for healthcare workers (topic 3) became one of the top third research topics. Besides, the appearances of new major human viral infectious always arise extensive research booms worldwide. Topic 4, 6, 11, 12, 18 that relating to the diagnosis and treatment of COVID-19 also developed rapidly. From July to August, the epidemic prevention and control situation in various countries became normalized. Around September, as some countries faced the second wave of the epidemic, the world COVID-19 epidemic showed a rebound. The development of research topics fell into the following four trends: (*a*) second increase after a brief stabilization: including topic 1, 8, 7, 10, 12, 16 and 22. These topics are much related to the dynamic development of the COVID-19 epidemic. With the continuous of the pandemic, these topics will ongoing. (*b*) stay stable: including topic 2, 4, 5, 6, 19, 20, 18, 23 and 24. These research topics have played an important role in the diagnosis, transmission, prevention and control of COVID-19. After long-term research and accumulation, certain research results and conclusions have been drawn. (*c*) Gradual downward: including topic 3, 15 and 13. The downward trend may due to the gradual maturity of related solutions and technical methods. (*d*) continued growth: including topic 9, 11, 14, 21 and 17. These topics relates to the application of emerging technologies of other research fields, such as deep learning and artificial intelligence, so their development was relatively late.

Conclusion

In this paper, we analyzed the monthly topic evolvements of COVID-19 related research using 30122 research papers' bibliometric data based on LDA model. We believe the results can help researchers and policy makers get deeper understanding of the development of the COVID-19 epidemic and the related research.

Acknowledgments

This work is supported by the Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2019176); Document Information Capacity Building Project of Chinese Academy of Sciences (No. E0290001); Project of National Science Library, Chinese Academy of Sciences (Optimization and Demonstration of Individualized Service of Document Information).

References

Al-Gheethi, A., Al-Sahari, M., Abdul Malek, M.; Noman, E., Al-Maqtari, Q., Mohamed, R., Talip, B. A., Alkhadher, S. & Hossain, M. S. (2020). Disinfection Methods and Survival of SARS-CoV-2 in the Environment and Contaminated Materials: A Bibliometric Analysis. *Sustainability*, 12, 7378.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

O'Callaghan, D., Greene, D., Carthy, J. & Cunningham, P. (2015). An Analysis of the Coherence of Descriptors in Topic Modeling. *Expert Systems with Applications*, 42(13), 5645-5657.

Are Non-Journal Citations More Important for Books? Study of Types of Citing Documents in Scopus

Ashraf Maleki

malekiashraf89@gmail.com
Independent Researcher, Tehran (Iran)

Introduction

Numerous studies have shown the fundamental contributions of books to humanities and social sciences in terms of the prevalence of publications (see Hicks, 1999) and the quantity of citations. Books either as research contributions, summary of previous research, educational materials and public and cultural communications have greater quantity of content than journal articles, and relatively the citations received and given are more numerous. For instance, books in Political Science tend to receive more numerous citations than the journal articles (Samuels, 2013), and thus represent important sources of impact in this field.

Despite a general difference between journal articles and books in terms of content and citation, it is less known whether journal citations are as useful as other sources of citations such as books and book chapters for book impact assessment. Formal citation indexing databases such as Book Citation Index (BKCI) and Scopus offer a relatively broad range of citations originating from various sources such as journal articles and in the past decade books and other resources as well. The main goal of this research is to investigate the difference between citations coming from these various document types in Scopus and examine their relationship with other books indicators such, Syllabus mentions as obviously an indicator of educational impact of books (Kousha and Thelwall, 2016), Goodreads Rating and Reviews (Zuccala et al., 2015) as an indicator of cultural impact of books, Library Holdings value of books (e.g. White et al., 2009), Google Books citations (Kousha and Thelwall, 2009) and Altmetric.com indicators. Therefore, this research is an attempt to analyse the types of documents citing books as identified by Scopus of Elsevier. This research is a follow-up investigation to previous research (Maleki, 2020) and aims to answer which types of citations originating from formal indexed document are statistically stronger sources of impact than journal articles for predicting citations to books and explaining other aspects of impact.

Method

In order to address research issue "cited by" feature in Scopus is used to harness the types of documents citing 36,493 books in six fields of Anthropology (1,198), Arts (1,245), Business and Economics (11,987), Law (4,734), Medicine (12,227), and Political Science (5,098). Other book indicators are extracted from Altmetric.com, Goodreads.com, Opensyllabus.org and Google Books, with API using ISBN, title and author names. Details of this data in terms of citedness in Scopus, Google Books (GB), Syllabus mentions (SM), Goodreads Users (GU), Goodreads Ratings (GR), Goodreads Text Reviews (GTR), Goodreads Average Rating (GAR), Library Print Holdings (LPH), Library E-holdings (LEH), and altmetrics indicators (Table 1) are offered in Maleki (2020).

Document Type in Scopus are provided in 18 different formats but the most important forms are Article, Book, Book Chapter, Conference Paper, Editorial, Letter, Note, and Review that are used to answer research questions. Citations originating from various document types are aggregated from Scopus Preview mode "cited by" results with URLs extracted via Scopus API in February 2021.

In response to research question, an Ordinary Least Squares Regression model is used on log-transformed data as proposed in Thelwall (2017) with citation counts across document types as independent variables and each other metric as the predicted one.

Findings and Discussions

Figure 1 indicates the proportion cited of books by each Scopus document type. It shows that more than 60% of books were cited by journal articles at least once in all fields, whereas for about 50% of books, book and book chapters citations was apparent, and about 40% review citations.

As the purpose of research is not to give a prediction formula but to examine the prediction ability of citations made by each document type only standardised coefficients of beta are reported instead of unstandardised B and the coefficients of only three document types out of nine document types entered in the model are reported as their coefficients was the strongest in the prediction of citation by every other citation source or predicted variable.

Results in Table 1 suggest that Book-sourced citations counts are often the best predictor of GB Citations, Syllabus Mentions, Library Print Holdings, and Goodreads related metrics, although the overall prediction are almost weak and never surpass 25.1% (for LPH). This shows that document types can only partially explain statistics of book impact although it would be better to prioritise and use Book-sourced citations instead of journal article citations.

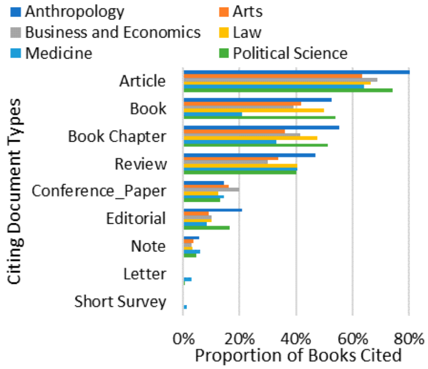


Figure 1. Proportion Cited of books in Document Type across six fields

Another interesting result in Table 1 is that articles-sourced citations to books are better for predicting Mendeley and Twitter and perhaps library electronic uptake of books than any other document type. Perhaps more interestingly is Review-sourced citations that dominate in the tiny prediction (1.4%-7.1%) made for News, Wikipedia and Blog citations to books, perhaps suggesting that reviews have a place in public communications with books.

Table 1. Statistically significant linear regression standardized coefficients (β) in six fields. The independent variables were all nine Scopus citing document types of which only three are reported for strongest coefficients.

Variables	*Article*	*Book*	*Review*	*Adjusted R2*
GB	.143	**.230**	-	19.9%
SM	.106	**.245**	-	21.7%
LPH	.220	**.311**	-.017*	25.1%
LEH	**.095**	.070	.060	4.1%
GU	.093	**.254**	.078	16.8%
GR	.045	**.249**	.088	15.2%
GTR	-.035	**.230**	.094	8.5%
GAR	.118	**.160**	.027	9.5%
Mendeley	**.232**	-.194	.108	9.9%
Twitter	**.156**	-.088	.090	3.6%
News	.075	-	**.098**	4.1%
Wikipedia	.042	-.029	**.105**	1.4%
Facebook	.073	**-.129**	.089	2.1%
Blogs	.092	.061	**.123**	7.1%

All coefficients are significant at p<0.001 except *p<0.05 and ** p<0.01. Bold coefficients represent the strongest document type as a source of citation in prediction of each metric. Negative coefficients are shown in red. Check abbreviations in Method.

Conclusions

This research provided a breakdown of document types of Scopus citations to books, in order to identify which formal source of citations helps in predicting various aspects of impact. Results indicated that although citing publication format is not a strong indicator of book impact there is some evidence to suggest that Book-sourced citation can be better role players in book impact assessment than journal-sourced citations when assessing traditional aspects of impact. However, Article-sourced citations sound to be strongly linked with online uptake of books such as in Mendeley and Twitter, indicating that online indicators mostly represent research value of books. Book Review citations might be linked to cultural value, however. In sum, citations from some document types tend to signal various aspects of impact which is worthy of attention.

References

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics, 44*(2), 193-215.

Kousha, K., & Thelwall, M. (2009). Google book search: Citation analysis for social science and the humanities. *Journal of the American Society for Information Science and Technology*, 60(8), 1537-1549.

Kousha, K., & Thelwall, M. (2016). An automatic method for assessing the teaching impact of books from online academic syllabi. *Journal of the Association for Information Science and Technology*, 67(12), 2993-3007.

Maleki, A. (2020). P-Libcitation vs. E-libcitation?: Libraries' Print Book Holdings Resonate with Citations and Altmetrics But E-book Holdings Do Not. *The 2020 Online Altmetrics Workshop*. Accessible at http://altmetrics.org/altmetrics20/

Samuels, D. (2013). Book Citations Count. *Ps-Political Science & Politics*, 46(4), 785-790. doi:10.1017/S1049096513001054

Thelwall, M. (2017). *Web indicators for research evaluation: A practical guide*. San Rafael, CA: Morgan & Claypool.

Zuccala, A. A., Verleysen, F., Cornacchia, R., & Engels, T. (2015). Altmetrics for the Humanities: Comparing Goodreads reader ratings with citations to history books. *Aslib Proceedings*, 67(3).

White, H. D, Boe, Yu. et. al. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology* , 60 (6), 1083-1096.

Identifying research themes in Finland using topic modeling and word embeddings

Katja Mankinen[1] and Yrjö Leino[2]

{[1] katja.mankinen, [2] yrjo.leino}@csc.fi
CSC - IT Center for Science, P.O. Box 405, 02101 Espoo (Finland)

Introduction

To identify research topics and trends in Finland in a fully data-driven and unsupervised way, we apply natural language processing and topic modeling methods to abstracts, titles, and keywords of research papers published in 2008-2019. Following the previous studies on topic models and mapping of Finnish science (e.g., Suominen & Toivanen, 2016), the present study enhances topic modeling methods by utilizing joint document and word embeddings to capture the semantic knowledge at document level. This work has been done in collaboration with the Academy of Finland in the context of Academy's State of Scientific Research analyses.

Data and methods

In total 106,736 publications (articles, reviews, letters, books) with at least one author with a Finnish affiliation, written in English and published in 2008-2019 were retrieved from Web of Science Core Collection. Only titles, abstracts and keywords of publications were considered for further analysis, excluding other metadata such as author and journal names as well as citations.

During data preprocessing, all words were tokenized, and stop words together with too rare and too common words as well as numbers and special characters were removed, after which unigrams and bigrams were extracted. To improve the quality of topics, the topic modeling method using the bag-of-words approach, Latent Dirichlet Allocation (Blei, Ng & Jordan, 2003), was extended to incorporate joint document and word semantic embeddings based on the model by Nguyen et al. (2015). The embeddings, created using doc2vec (Le & Mikolov, 2014), can better describe the connections between words and documents, and were used to form interpretable topic representations summarizing common words of related documents. Finally, each publication was assigned to the most significant topic.

To evaluate the research impact in these newly identified research topics, the top 10% citation index (proportion of publications belonging to the top 10% most frequently cited publications, with world average of 1.0 (Waltman & Schreiber, 2013) was calculated for each topic.

Results

In total, 1026 research topics were found. Number of publications in each topic varies between 18 and 862 with the mean being 104.0 and median 81.5. Topics with the most publications represent broad disciplines from education and business management to wireless networks, renewable energy and gut microbiota, whereas the smallest topics tend to be very specific such as distinct animal or plant species (e.g. frogs, spruces) or diseases (e.g. carpal tunnel syndrome). In addition, a few small low-quality categories were found, in which no clear connections can be found between more than a couple of words.

Topics with high scientific impact

We found 65 topics with high scientific impact (the top 10 index over 2.0) and 11 topics with very high scientific impact (the top 10 index over 3.5). Examples of these topics from various research disciplines are shown in Table 1.

Table 1. Examples of topics labelled using their most significant keywords, the top 10 index and full count of publications.

Topic	Top 10 index	Publications
RFID technologies	3.59	172
Wireless networks	3.23	801
Arctic maritime	2.91	197
Laser scanning, remote sensing	2.90	388
(Nano)cellulose	2.75	324
Permanent magnet motors	2.68	171
Ecology, climate change and biodiversity	2.25	210
Optimization	2.18	168
Sports and exercise medicine	2.10	102
Customer value, servitization	2.06	352

A majority of the topics with the highest top 10 indices is related to various aspects of computer science, electrical engineering, environmental sciences, materials science, telecommunications and management.

Dynamics of topics

Dynamic changes in the publication volume over time were explored in order to observe emerging research trends. Examples of topics with increasing number of publications are presented in Figure 1. The figure shows that there has been a steady increase in the number of publications especially in renewable energy and education research in recent years. No sharply peaking topics were found, and none of the topics showed a rapid decrease in the number of publications in 2008-2019.

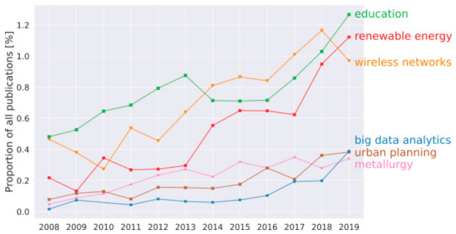


Figure 1. Relative publication volume in selected topics in 2008-2019 normalized to yearly count of publications to account for general increase in publication volume.

National and international collaboration

For each topic, we studied national and international co-authored publications and evaluated their scientific impact. In some topics, such as in “wireless networks”, international collaboration is emphasized while national collaboration is almost an exception. In many other topics, such as in “laser scanning and remote sensing”, national collaboration is extensive and has mostly high impact. Scientific impact in topics can also be strongly polarized: of all organizations active in a topic, maybe only few are doing research on an exceptional level. Therefore, high national impact does not directly imply that impactful research is conducted throughout Finland.

Discussion

A majority of resulting topics represent easily interpretable, large-scale research phenomena. The same general theme can be addressed in many topics from different perspectives. For example, there are many topics related to oncology, usually distinguished by an organ of interest (“prostate cancer”, “cervical cancer”). Topics of global interest with distinct vocabulary, such as one about supernovae, tend to contain publications only from one or two disciplines (astronomy). On the other hand, there are topics with only loosely connected publications from several separate research fields: e.g. a topic “music” covers a wide range of research from emotions (psychology) to brain responses (neurosciences) and from music theory to pedagogy. This also reflects differences between disciplines: some disciplines are inherently further away from others, while others often partly overlap, such as biomedicine, medical and health sciences.

It should be noted that not all publications in the topic are necessarily strictly related to the topic, and not every publication concerning the topic in question has been clustered into it. Even significant topics may remain unrecognized due to limitations in data (only a subset of all Finnish research published in 2008-2019, leading to low coverage especially in humanities and social sciences; occasional short or missing abstracts) and methods (model parameter choices, text preprocessing). While used methods improved the interpretability of topics and are suitable for mapping research themes at scale, they are not suited for accurate classifications or determining the exact size of research areas. Our ongoing work will extend this study by making use of citation information and large, pretrained neural network models.

Conclusions

We recognized several – perhaps even a few previously overlooked – research topics where high-impact research is done in Finland. Combined with a detailed analysis of research collaboration, this study provides data-driven insights of the Finnish research field.

Acknowledgments

Results presented here are derived from the Web of Science® prepared by CLARIVATE ANALYTICS ®, Inc. (Formerly the IP & Science business of Thomson Reuters®), Philadelphia, Pennsylvania, USA: © Copyright CLARIVATE ANALYTICS ® 2021. All rights reserved. Results are taken with permission from the bibliometric analysis system provided by CSC - IT Center for Science Ltd., Espoo, Finland.

References

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Le, Q. & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31th International Conference on Machine Learning* (pp. 1188-1196).

Nguyen, D., Billingsley, R., Du, L. & Johnson, M. (2015). Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics*, 3, 299-313.

Suominen, A. & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, 67(10), 2464-2476.

Waltman, L. & Schreiber, M. (2013), On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, 64(2), 372-379.

International collaboration of post-Soviet countries: what has changed in post-Soviet period?

Nataliya Matveeva[1], Andrey Lovakov[2] and Ivan Sterligov[3]

[1] *nmatveeva@hse.ru*, [2] *lovakov@gmail.com*
HSE University, 11 Pokrovsky Bulvar, 109028, Moscow (Russia)

[3] *ivan.sterligov@gmail.com*
HSE University, 11 Pokrovsky Bulvar, 109028 Moscow (Russia)
Center for Sociological Research, Potapovsky pereulok, 5, 101000 Moscow (Russia)

Introduction

Structure of academic research in a particular country is directly linked with the social, economic and political institutions that exist in that country (Graham, 1993). One of the major characteristics of such structure nowadays is the role and scope of international collaboration (Sonnenwald, 2007; Leahey, 2016). Development of science in particular countries can be significantly determined by the type of scientific activity with the other countries (Lauk & Allik, 2018).

In presented work we analyse international collaboration of post-soviet countries in dynamics. Post-Soviet region has several features, which make it a unique unit for analysis of the scientific structure and collaboration (Krementsov, 1996).

Thus, the common past and diverging present make the task of measuring and interpreting international collaboration for ex-USSR very interesting. In this study we address the following questions: Which patterns of international collaboration have emerged and developed in the ex-USSR during the post-Soviet era? What are the changing roles of Russia, EU and other major international forces in this collaboration?

Data and methods

Our sample consists of 15 post-soviet countries. We used data about the total number of journal articles and reviews in 1993, 1998, 2003, 2008, 2013 and 2018. Data was attributed to countries' profiles in WoS (indexes SCI-expanded, SSCI and A&HCI, document types “article” and “review”). We also use data about the number of publications in Q1 and Q4 journals according to their Journal Impact Factor. Our dataset does not cover many local journals including Russian-language venues, which is popular in former USSR. In this work we focus on internationally visible academic output using a well-researched and curated data source (Birkle et al., 2020).

For each country we analyse the total number of publications, number of publications in Q1 and Q4 segments and different research areas, the number of publications written in collaboration with other countries and without them. For the years 2010-2018 we are also able to study precise national composition of authors of collaborative papers using fractional counting (for earlier period this is impossible due to WoS limitations).

Findings

In the observed period the share of international collaboration of post-Soviet countries has increased significantly (Figure 1). All countries have higher values than average global baseline (black curve). In 1993, for different countries the value of this parameter was from 10% to 50%. In 2019 it varies from 40% to 100%. Russia has the lowest share of international collaboration that can be explained by the size of this country, which conforms to global trends of smaller countries being more international (Kamalski, 2009).

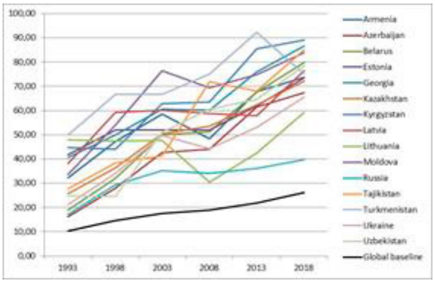


Figure 1. Share of international collaboration of post-Soviet countries.

The highest values of international collaboration are observed in Turkmenistan and Kyrgyzstan, which have the lowest total WoS paper counts. Lithuania is the most stable country by share of international collaboration: during the period it has changed from 47.9% to 59.7%.

International collaboration of post-Soviet countries also varies by disciplines. In 1993, post-Soviet countries most often collaborated with other countries in Natural Science and Engineering & Technology. In 2018, some countries intensified international collaboration in Humanities fields. During the observed period the share of international collaboration in Agricultural Sciences has increased almost tenfold.

We also observe that international collaboration of post-Soviet countries in the Q1 segment is much higher than in Q4. Moreover, in 2018 for all countries in Q1 segment share of international collaboration is more than 70%. Thus, international collaboration is more pronounced in higher-cited journals.

After the collapse of the Soviet Union on average post-Soviet countries collaborate with each other in the same proportion as with other countries (Figure 2). Then the share of collaboration with other countries has increased. In 2018 papers written by post-soviet countries in collaboration with each other constitute 10% from all international publications.

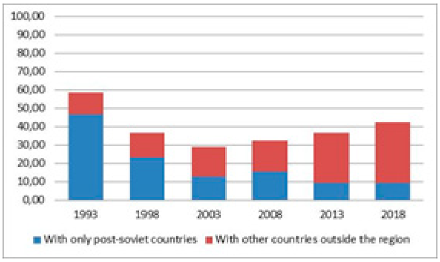


Figure 2. Average share of papers written with countries inside and outside the region.

In 1993 for all post-Soviet countries Russia was the main collaborator. In 2018, Russia isn't the main collaborator for Baltic countries and Georgia. Currently, the main collaborators of post-Soviet countries are also the USA and Germany.

We conduct fractional counting of countries' shares for all Russian papers with all ex-USSR countries published in 2009-2019 and having 2-20 authors (thus we intentionally exclude large-scale international projects and megascience). Such analysis shows that the ratio of Russian to local authors in collaborative papers is almost always less than 1, with the exception of Kazakhstan and Lithuania (Figure 3).

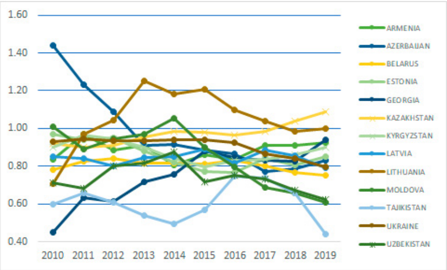


Figure 3. Share of Russian and local authors in collaborative Russian+exUSSR papers with 2-20 authors, 3-year moving average.

Role of non-exUSSR countries in post-Soviet collaborations is very different for various countries, but tends to increase. It varies from 15% for Kazakhstan and Belarus to 30% for Baltic states, Azerbaijan and Georgia.

Discussions and conclusions

Our preliminary analysis shows that after the collapse of Soviet Union post-Soviet countries significantly changed the patterns of international collaboration, and these changes were country-specific. We observe a dramatic decrease of scientific collaboration between post-Soviet countries. While the relative role of Russia as a partner is decreasing for all the ex-USSR nations, this dynamic differs a lot, and tends to be aligned with broader geopolitical and foreign affairs agenda.

Acknowledgments

This work was supported by the Russian Science Foundation (grant 20-18-00140).

References

Birkle, C., Pendlebury, D. A., Schnell, J. & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. Quantitative Science Studies, 1(1), 363-376.

Graham, L. B. (1993). *Science in Russia and the Soviet Union: A Short History*. Cambridge University Press.

Kamalski, J. (2009). *Small countries lead international collaboration*. Elsevier Research Trends 14. https://www.researchtrends.com/issue14-december-2009/country/

Krementsov, N. (1996). *Stalinist science*. Princeton University Press.

Lauk, K. & Allik, J. (2018). A Puzzle of Estonian Science: How to Explain Unexpected Rise of the Scientific Impact. *Trames. Journal of the Humanities and Social Sciences*, 22(4), 329-344.

Leahey, E. (2016). From sole investigator to team scientist: Trends in the practice and study of research collaboration. *Annual review of sociology*, 42, 81-100.

Sonnenwald, D. H. (2007). Scientific collaboration. Annual Review of Information Science and Technology, 41, 643-681.

Building Multi-level Aspects of Peer Reviews for Academic Articles

Minghui Meng[1], Yuzhuo Wang[2] and Chengzhi Zhang[3]

{[1]mengmh, [2] wangyz, [3]zhangcz}@njust.edu.cn
Nanjing University of Science and Technology, 200 Xiaolingwei Street, Nanjing 210094 (China)

Introduction

Peer reviews for academic articles are domain experts' comments on the paper submitted to a journal or conference, which reflect the overall impression of the reviewer as well as the detailed comments.

By mining peer reviews at aspect level, we can find the aspects concerned by reviewers, which can provide direction for inexperienced paper submitters to optimize their writing. Some scholars have studied it at aspect level, Chakraborty et al. (2020) adopted the reviewing aspects used in the ACL conferences, including *clarity, originality,* etc. However, these aspects are the overall impression of the reviewer on the paper with coarse granularity and lack of more detailed comments. Moreover, reviewers of different disciplines may focus on different aspects.

Therefore, we want to find more fine-grained multi-level aspects from multi-disciplinary peer reviews, including domain independent aspects and domain related aspects, which can provide a well-rounded optimization direction for submitters. We take peer reviews in *Nature Communications* as an example and propose a general method for building multi-level aspects of peer reviews for academic articles.

Method

The process of this paper is shown in figure 1:

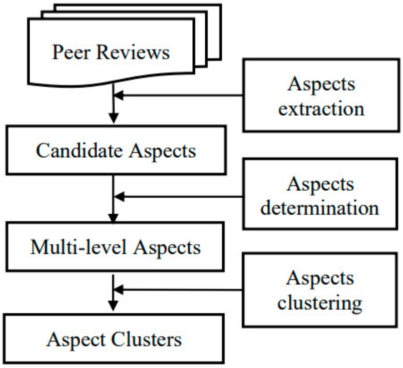


Figure 1. Flowchart of this study

Data collection. *Nature Communications*[i](NC) is a multidisciplinary journal, which has 5 first-level disciplines and 71 second-level disciplines. Besides, we regard "NC" as the zero-level discipline. It has provided the access to content of peer review since 2016. We collect 187,971 peer review documents of articles published from 2016 to 2020, details of the NC's first-level disciplines are shown in table 1:

Table 1. Distribution of NC Peer Reviews

Discipline	Papers	Reviews
Biological sciences	21,683	98,585
Earth and environmental sciences	2,588	11,632
Health sciences	6,090	28,674
Physical sciences	10,180	45,222
Scientific community and society	832	3,858

Candidate aspects extraction. We use the Double Propagation algorithm (Qiu et. al., 2011) to extract aspects from peer reviews. Firstly, we employ the StanfordNLP[ii] tool to tag the part of speech of peer reviews and parse sentences. Secondly, we select words in opinion lexicon from Liu [iii] as seed opinions. Next, we consider aspects to be nouns and opinions to be adjectives and limit the dependency relations between aspects and opinions to mod, subj, etc. Finally, we select all the words matching the above rules as the candidate aspects.

Multi-level aspects determination. We divide multi-level aspects into domain-independent common aspects (DICA), domain-related common aspects (DRCA), and domain-related special aspects (DRSA). Multi-level aspects are determined based on the number of reviews, the distribution uniformity, and particularity of candidate aspects at all levels of domain peer reviews. We use the inter-domain entropy (IDE) and termhood (Chang, 2005) to measure the distribution uniformity and particularity of aspects in various domains, the calculation formulas are as follows:

$$\mathrm{IDE}(w_i) = -\sum_j P_{ij} \log P_{ij} \qquad (1)$$

$$\mathrm{Termhood}(w_{ij}) = n_{ij} \times log_2 \left[\frac{N}{Nd_i}\right] \qquad (2)$$

Where P_{ij} is the probability of candidate words w_i in domain J,$Nd_i = 2^{IDE(w_i)}$, N is the total number of the disciplinary domains.

The object of this study is peer reviews with three-level disciplinary domains, like NC. The following is the process to determine the multi-level aspects of NC, including DICA, DRCA and DRSA. DICA is the aspect that often appears and distributes evenly in every first-level domains. DRCA often appears in a certain first-level disciplinary domain and is evenly distributed in its sub-domains. DRSA is the aspect unevenly distributed in each second-level domains and often appears in a certain second-level domain. First, we count the number of peer reviews where the zero-level candidate aspects at NC's first-level domains and calculate the IDE, next select words with the top 80% cumulative percentage of total word frequency, then we get the words with entropy higher than median as the final DICA. Second, we count the number of peer reviews where NC's first-level candidate aspects at corresponding second-level domains and calculate the IDE, then calculate its termhood at each second-level domain, next remove DICA, according to the same filtering conditions, get the words that meet the conditions as DRCA. After filtering DICA and the corresponding DRCA, we obtain the top100 termhood words of each second-level domain as the final DRSA.

Aspects clustering. In this paper, we select the Affinity Propagation (AP) algorithm (Frey & Dueck, 2007) to cluster aspects, the purpose is to aggregate the same aspects of different expressions. We train word2vec[iv] model to represent word vector, then use cosine similarity to calculate the similarity between two aspects and choose the Silhouette Coefficient (SC) to evaluate the clustering result.

Results

The extraction result of candidate aspects

We respectively extract the candidate aspects from NC's every-level discipline peer reviews. As a result, we extract 5896 zero-level candidate aspects, 5 groups of first-level candidate aspects, and 71 groups of second-level candidate aspects.

The determination result of multi-level aspects

In NC peer reviews, we obtain 598 DICA, 5 groups of DRCA, and 71 groups of DRSA. Table2 is the partial result of the NC multi-level aspects, including NC's DICA, Biological sciences (BIO)'s DRCA and Computational biology and bioinformatics (CBB)'s DRSA.

Table 2. Multi-Level Aspects

Level	Category	Domain	Aspects
0	DICA	NC	Data\result\method\figure\analysis\change\experiment\effect\text\level\model\topic\example\...\discussion\
1	DRCA	BIO	Assay\target\domain\...\specificity
2	DRSA	CBB	Object\update\classifier\...\paradigm

The clustering result of zero-level discipline aspects

In AP clustering, the preference value refers to the reference degree of the point as the clustering center, and the damping factor is the coefficient used for convergence. We cluster NC's DICA. Our clustering requirement is to control the number of clusters between 10 and 50, so that clusters have both clustering effect and discrimination, we set preference from -50 to -25, and damping from 0.5 to 0.8. When preference = -40, damping = 0.65, the SC value was the largest, we regard it as the optimal clustering result. Table3 is part of the clustering results of NC's DICA. There are four clusters, which are "Result", "Impact", "Method", and "Figure". In the "Result" cluster, reviewers pay more attention to *result, conclusion,* and *finding*. In the "Impact" cluster, reviewers focus on *impact* and *lack*. Besides, *method*, *figure*, etc are also the focus of reviewers.

Table 3. The NC's DICA Clusters

Cluster	Aspects	# Member
Result	Result\conclusion\finding\...\claim	19
Impact	Impact\lack\contribution\...\focus	48
Method	Method\approach\technique\...\strategy	19
Figure	Figure\image \table\legend \...\panel	29

Conclusion

This paper proposes a method for building multi-level aspects of peer reviews for academic articles, which aims to comprehensively extract the multi-level aspects in peer reviews. The multi-level aspects reflect the key aspects that reviewers focus on, such as *result* and *method*, which can provide reference for inexperienced submitters to optimize their research design and writing.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 72074113) and Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. SJCX20_0137).

References

Chakraborty, S., Goyal, P. & Mukherjee, A. (2020). Aspect-based Sentiment Analysis of Scientific Reviews. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20)* (pp.63-81). New York, USA: ACM.

Qiu, G., Liu, B., Bu, J. & Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1), 9-27.

Chang, J. (2005). Domain Specific Word Extraction from Hierarchical Web Documents: a First Step toward Building Lexicon Trees from Web Corpora. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing* (pp.64-71). Stroudsburg, PA: ACL.

Frey, B. J. & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315(5814), 972-976.

[i] https://www.nature.com/ncomms/

[ii] http://nlp.stanford.edu/software/tagger.shtml.

[iii] http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

[iv] http://word2vec.googlecode.com/

Multilingualism of Ukrainian Humanities: how it is seen globally

Serhii Nazarovets[1] and Olesya Mryglod[2]

[1] *sergiy.nazarovets@gmail.com*
State Scientific and Technical Library of Ukraine, Antonovycha Str 180, 03680 Kyiv (Ukraine)

[2] *olesya@icmp.lviv.ua*
Institute for Condensed Matter Physics of the National Academy of Sciences of Ukraine, Svientsitskii Str 1,
79011 Lviv (Ukraine)

Introduction

Dissemination of knowledge in the Humanities occurs through the use of different languages and different types of documents. Many studies of recent years consider the issue of proper evaluation of scientific non-English achievements of Humanities (Pedersen, Grønvad, & Hvidtfeldt, 2020). Such a lively interest can be partially explained by many examples of misuse of metrics in order to assess the impact of research in Humanities. In recent years such popular initiatives as DORA, Leiden Manifesto, Metric Tide, Helsinki initiative attracted the attention of the scientific community to the problem of correct evaluation of non-English works. However, the other side of the coin, which is related to the importance of international evaluation of results in Humanities, is not always properly addressed to by researchers. The exclusive focus on the local level also leads to negative consequences: lack of independent review, decrease of the quality and relevance of scientific results.

The research policy of Ukraine, which became independent in 1991, inherited many Soviet remnants. For a long time scientists in Ukraine have been rewarded primarily for publications in national journals in their native language (Hladchenko, & Moed, 2021). The locally-nested character of research together with such formal stimulations create the preconditions for harmful self-isolation and "invisibility" of Ukrainian Humanities.

The aim of this work is to investigate the volume and language distribution of publications by Ukrainian researchers in Humanities indexed in Web of Science (WoS). Further, a number of other non-English-speaking countries of Eastern Europe with similar post-Soviet background is chosen for comparison.

Language is an important factor influencing the coverage of a country's scientific publications by Web of Science, especially in the Humanities (Mongeon & Paul-Hus, 2016) for non-English-speaking countries. However, this database – one of the most used for scientometric studies – can be considered as a prism through which national research output is seen globally. Therefore, despite all cautions, the data from Web of Science Core Collection are used in this work.

Method and Data

The metadata about publications by Ukrainian authors (at least one Ukrainian affiliation per paper) in Philosophy, History and Literature in 2010-2019 from WoS Core Collection were used: total number of records, language distribution, information about journals where English-language works were published. The comparative analysis was performed using the same data (same period) for few other non-English speaking Eastern European Slavic countries: Poland, Czech Republic and Slovakia. An example of search query: CU=Ukraine AND PY= (2019 OR 2018 OR 2017 OR 2016 OR 2015 OR 2014 OR 2013 OR 2012 OR 2011 OR 2010) AND SU=Philosophy. The date of data retrieval is 10.01.2021.

Results

Our results show that according to WoS, choosing different languages for publications is typical for scientists from the analyzed countries of Eastern Europe in the fields of Philosophy, History and Literature (see Fig. 1-3). The share of publications in other than national language is explicitly defined if it is significant: Russian language for Ukraine; Slovak language for Czech Republic; and Czech language for Slovakia.

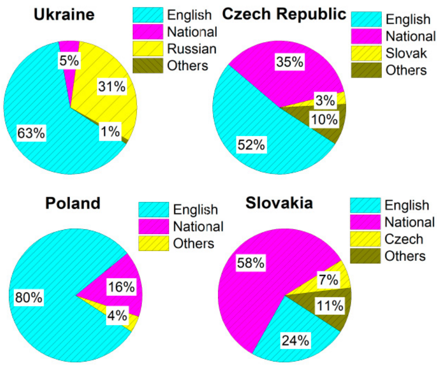


Figure 1. Language distributions of publications by authors from different countries: Philosophy (2010-2019).

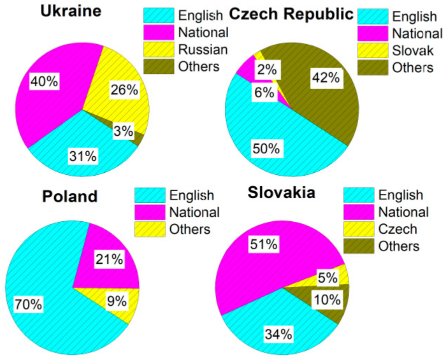


Figure 2. Language distributions of publications by authors from different countries: History (2010-2019).

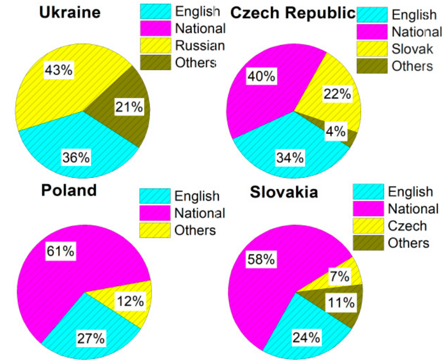


Figure 3. Language distributions of publications by authors from different countries: Literature (2010-2019).

Depending on the discipline and country, the ratio of English-language publications in other languages varies. For example, the shares of papers in the field of Philosophy written in English by Polish and Slovak authors during 2010-2019 are significantly different: 81% and 24%, correspondingly. At the same time, in the fields of History and Literature, scholars from all considered countries demonstrate similar behavior in the context of publication language choice.

In addition, our results shows that a significant share of English-language papers in Philosophy, History and Literature by authors from Eastern European countries are found in the journals published in the same country.

Conclusions

The results confirm the findings of previous studies: it is natural for Humanities to publish research outputs in various languages not being concentrated exclusively on English. This is quantitatively illustrated for Ukrainian, Polish, Slovak and Czech authors. Moreover, language particularities are observed for different countries. Due to historical circumstances, Russian language plays substantial role for Ukrainian Humanities, which is particularly notable and rather counterintuitive for the Literature. But only since 2017 Ukrainian becomes the official language of the educational process. E.g., Russian language is much less influential in Poland, where it was compulsory until 1989, see (Kulczycki, Rozkosz, & Drabek, 2019). Similarly, the mutual language influences are noticeable for scholar publications of Slovak and Czech Republics.

The majority of English-language articles in Humanities by researchers from the considered Eastern European countries and indexed in WoS correspond to the journals published in the country of authors. This practice needs to be studied more carefully, but there is an agreement with conclusions drawn in (Nazarovets, 2020).

Unfortunately, the lack of sufficiently large statistics does not allow to perform the temporal analysis of Ukrainian publication data presented in WoS. The potential changes in language spectrum in response to implementation of new rules would provide an answer to the question about the possible impact of research policy on the scholar publication strategy.

Our preliminary results suggest that natural multilingualism of Humanities has to be taken into account during the process of research evaluations, especially based on quantitative approaches and metrics.

References

Pedersen, D. B., Grønvad, J. F. & Hvidtfeldt, R. (2020). Methods for mapping the impact of social sciences and humanities—A literature review. *Research Evaluation*, 29(1), 4-21. https://doi.org/10.1093/reseval/rvz033

Hladchenko, M. & Moed, H. F. (2021). National orientation of Ukrainian journals: means-ends decoupling in a semi-peripheral state. *Scientometrics*, 126(3), 2365-2389. https://doi.org/10.1007/s11192-020-03844-4

Kulczycki, E., Rozkosz, E. A. & Drabek, A. (2019). Internationalization of Polish Journals in the Social Sciences and Humanities: Transformative Role of The Research Evaluation System. *Canadian Journal of Sociology*, 44(1), 9-38. https://doi.org/10.29173/cjs28794

Mongeon, P. & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106, 213-228. https://doi.org/10.1007/s11192-015-1765-5

Nazarovets, S. (2020). Controversial practice of rewarding for publications in national journals. *Scientometrics*, 124(1), 813-818. https://doi.org/10.1007/s11192-020-03485-7

Rigor and Transparency Index for Systematic Literature Reviews: a first stage approach

Gerson Pech[1] and Catarina Delgado[2]

[1]*pech@uerj.br*
Department of Nuclear Physics and High Energies, Rio de Janeiro State University, 20550-900 Rio de Janeiro, (Brazil)

[2]*cdelgado@fep.up.pt*
Faculty of Economics, University of Porto, 4200-464 Porto, (Portugal)
LIAAD/ INESC TEC, University of Porto, 4200-465 Porto, (Portugal)

Introduction

Systematic literature reviews (SLRs) play a crucial role in science as they are responsible for systematizing the main results of the key questions of different research fields (Pech & Delgado, 2021). The quality of these works is related to the development of an appropriate review protocol, to the impartiality in papers' selection, and to the rigor in analyzing the data of the studies used. In this way, the AMSTAR**,** AMSTAR 2 (Shea et al., 2017), and PRISMA (Moher et al., 2009) guidelines have been applied in many SLRs as the main methodological supports for their development, and the evaluation of their quality. However, although some research has been carried out to assess the adherence level of the review studies to these two guidelines, none, so far, has carried out a large-scale longitudinal assessment covering different research areas. That is the issue addressed by this work. The objective of this study is to construct the basis of a tool able to, automatically, measuring the rigor and transparency level of SLRs. In fact, the development of a metric to assess the quality of reviews on a large scale based on the methodology would be important to understand their temporal evolution and their dependence in terms of different fields. This approach was used by Menke et al. (2020), not for SLRs, but to evaluate Medicine and Biology papers. The authors created an automated tool (sciscore.com) to assess the quality, based on rigor and transparency of the methodologies used in research of these fields. For this, they created the Rigor and Transparency Index (RTI) as a measurement of the method adherence used in these studies to the established guidelines. Considering the metric created by the authors, the higher the paper's RTI, the higher the guarantee that the study can be reproduced by other researchers. The same concept was applied in this paper, but for SLRs. This is a work in development, and in this first stage, we defined the RTI for Systematic Literature Review (RTI-SLR) using only the AMSTAR and PRISMA items related to the methodology and results' presentation. For each item, we generated lists of key-terms that were used to calculate the RTI-SLR in 100 selected articles.

Method

Our method consists of four steps, as shown in Figure 1.

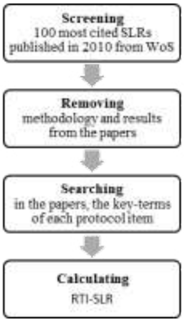

Screening
100 most cited SLRs published in 2010 from WoS
Removing
methodology and results from the papers
Searching
in the papers, the key-terms of each protocol item
Calculating
RTI-SLR

Figure 1. The method used in this study

In the first step, we retrieved the 100 most cited articles containing the expressions "systematic literature review" OR "systematic review" in TITLE-ABS-KEY and published in 2010. We used the Web of Science database, as it is widely used in many research fields. The extraction was carried out on January 3, 2021. We used papers published in 2010 since the exposure time of these papers was at least 10 years. This made it possible to compare the number of citations of the paper with the RTI-SLR value. In the second step, we extracted the sections in which the methodology and results of the review study are presented. Here, we also used the results section, instead of only the methodology section, as used by Menke et al. (2020). The third step is designated to search the key-terms in the methodology and result sections. Table 1 shows the list of some key-terms classified in the eight methodological items defined in this work (hereafter, protocol-item). This search was done with the NVIVO 1.3, using the words' frequency function. In this way, we applied NVIVO to calculate the occurrence of the key-terms of each protocol-item in all the papers included in our database.

Table 1. Key-term of each protocol-item

Protocol-item	*Key-terms*
Eligibility criteria (EC)	“search strategy”; “inclusion criteria"; “exclusion criteria”
Duplicate study selection (DS)	“independently reviewed”; “independently by two researchers”; “reviewers independently”
Data collection process (DC)	“selection process”; “process of study selection”; “screening process”; “search process”; “review process”
Different databases (DD)	“Scopus”; “Web of Science”; “WoS”; “PubMed”; “Medline”; “Embase”; “Google Scholar”
Search terms (ST)	“string of search”; “search query”; “literature search”; “literature selection”
Flow chart of the process (FC)	“study flow”; “flow diagram”; “flow of articles”; “PRISMA”; “AMSTAR”
Status of the publications (SP)	“publication status”; “language”; “published in English”; “published from”; “published between”; “English-language”
Characteristics of the papers (CP)	“characteristics of studies”; “description of studies”

Finally, the last step calculates the RTI-SLR$_i$ for each paper i, using the following equation:

$$RTI\text{-}SLR_i = \sum_{\alpha=1}^{\alpha=8} Q(N_{i\alpha}) \quad (1)$$

where α defines the protocol-item, $N_{i\alpha}$ is the frequency which the key-terms of the protocol-item α were found in article i, and $Q(N_{i\alpha})$ refers to the quartile in which $N_{i\alpha}$ is found considering all the papers analyzed in this study. $Q(N_{i\alpha})$ is equal to 4, if $N_{i\alpha}$ is from the first quartile; 3, if it is from the second quartile, and so on.

Results

Figure 2 shows the WoS categories of the sample. We only represented categories with more than two papers. All the sample papers are from only one WoS Research Areas: Life Sciences & Biomedicine.

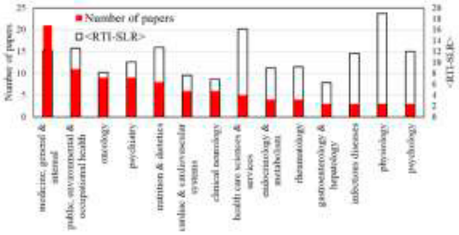


Figure 2. WoS categories of the papers. The red bars denote the number of papers. The open bars represent the mean RTI-SLR whose values are on the right axis.

In this sample, the journal that has more papers is The Lancet**.** Figure 2 also shows the average RTI-SLR value (<RTI-SLR>), indicated on the right vertical axis, for the sample papers belonging to each of the WoS categories. The value does not depend on the number of articles in the category; however, they vary from 6 (critical care medicine) to 19 (physiology) for this set shown in the figure, showing that in several areas the reproducibility of the research may be higher.

Conclusions and perspectives

This paper presented a method to assess the rigor and transparency of the methodologies that have been applied in the SLRs. The results showed that the key-terms used in the model have a close relationship with the process establish stated by the PRISMA and AMSTAR guidelines. Besides, our approach, specifically, the assigned of relevant protocol-items, and the selection of appropriate key-terms for each protocol-item may be a perspective of the first stage for the RTI-SLR development. This initial study suggests that, with further developments and analysis of this tool, it will be possible to extend this model to apply it on a large scale, to assess the rigor and transparency level of SLRs of different periods and fields. In this way, future works should address, mainly, the following improvements: (i) including other methodological items that must be evaluated; (ii) complete the list of key-terms for each protocol-item; (iii) elaborating the large-scale analysis longitudinally; (iv) measures the correlation between the mean RTI-SLR and the Journal Impact Factors; and (v) comparing the RTI-SLR scores with other research areas that usually develop SLRs, such as management and Physics.

References

Menke, J., Roelandse, M., Ozyurt, B., Martone, M., & Bandrowski, A. (2020). The Rigor and Transparency Index. Quality Metric for Assessing Biological and Medical Science Methods. *iScience, 23*(11):101698.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G., (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Bmj-British Medical Journal, 62*, 1006-1012.

Pech, G., & Delgado, C. (2021). Screening the most highly cited papers in longitudinal bibliometric studies and systematic literature reviews of a research field or journal: Widespread used metrics vs a percentile citation-based approach, *Journal of Informetrics,15*(3), 101161.

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., et al. (2017). AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non- randomised studies of healthcare interventions, or both. *Bmj-British Medical Journal*, 358:j4008.

Users of National Open Access Journals: The case of Journal.fi platform

Janne Pölönen[1], Sami Syrjämäki[2], Antti-Jussi Nygård[3] and Björn Hammarfelt[4]

{[1] janne.polonen, [2] sami.syrjamaki@tsv.fi, [3] antti-jussi.nygard}@tsv.fi
Federation of Finnish Learned Societies, Snellmaninkatu 13, 00170 Helsinki (Finland)

[4] bjorn.hammarfelt@hb.se
Swedish School of Library and Information Science, University of Borås, Allégatan 1, Borås, 503 32 (Sweden)

Introduction

National language publishing and open access (OA) are claimed to facilitate dissemination of research knowledge beyond academia (Late et al, 2020; Zuccala, 2009), however their contribution to the impact of research remains a blind-spot.

Citation analysis based on Web of Science and Scopus databases provides only a partial picture of research impact (Nicholas et al., 2005), partly because they cover mainly English language articles (Kulczycki et al, 2020. Altmetrics aim to capture the wider outreach of research, and in this study, we focus on "use" in terms of a request for a particular source (cf. Kurtz & Bollen, 2010).

The most promising altmetric sources for user profiles, such as Mendeley, mainly cover research published in English language journals (Wouters & Costas, 2012; Haustein, 2014; Mohammadi et al., 2015; Hammarfelt, 2014). However, OA journal platforms, such as the Érudit platform (Cameron-Pesant, 2019), the Croatian open access platform Hrcâk (Stojanovski, Petrak, & Macan 2009) or the Journal.fi platform studied here offer the possibility to analyse the use of journals, which are published in languages other than English.

Research questions, data and methods

In this paper we study the diversity of users of OA articles on the Finnish Journal.fi platform. This platform hosts 98 OA journals (spring 2020) from all fields of arts and sciences publishing in different languages, mainly Finnish and English. Our main research questions are:

1. Who are the users of articles published on Journal.fi platform? In what role they are using articles from journal.fi platform and what is their geographical distribution?

2. Which kinds of publications are the different groups of users interested in? Do their interests differ according to the year and language of publication?

Federation of Finnish Learned Societies planned an open online survey to visitors of article abstracts and full-texts on the Journal.fi platform. 48 journals (50% of all journals) participated in the study, and the survey was active from 7 February until 31 March 2020.

The survey was organized by using a plugin created for the Open Journal Systems platform. Each visitor was presented in a pop-up window an invitation to join the survey, in which they were asked to indicate one role in which they read or search Journal.fi articles from a list of choices. Participants also permitted tracking cookies for storing data about the Journal.fi articles and abstracts they visited during the survey. Each visitor was identified with a unique hash key and the visitors' IP address was used to determine a geolocation. Information about the year and language of publication of articles was gathered from the Journal.fi platform.

Results

Total of 668 users participated in the survey. The two largest groups were students (40%) and researchers (36%), followed by private citizens (8%), other experts (7%) and teachers (5%). Other identified user roles include journalists, civil servants, entrepreneurs and politicians. Students (42%) are clearly the largest user group of national language (Finnish and Swedish) publications, and besides researchers (25%), also private citizens (12%) and other experts (11%) figure prominently among users. For the foreign language publications, researchers (46%) and students (38%) are more clearly the main user groups.

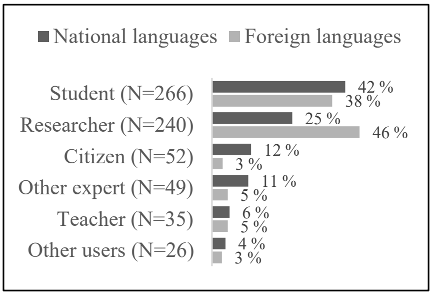


Figure 1. Share of 668 users of national and foreign language articles by self-reported role

The survey participants visited 1,546 articles a total of 2,018 times (counting only one visit by participant per article). As shown in Figure 2, article visits by students are less focused on the latest articles compared to other groups. The vast majority, 97%, of visits to national language articles are by users from Finland. The foreign (mainly English) language articles serve a much more international audience: 37% of the article visits are by Finnish users, 5% come from the Nordic countries and 58% the rest of the world.

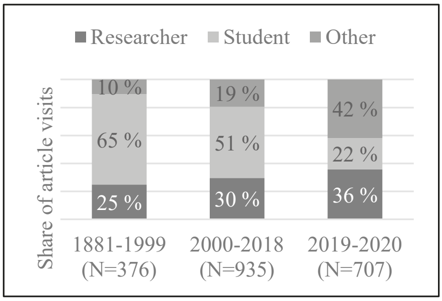


Figure 2. Share of 2,018 article visits by different user groups and year of publication.

Discussion and conclusions

Our study shows that OA publications in national languages are vital for reaching important users of research both within and beyond academia, including experts, citizens, teachers and students. In the specific context of Finland, these groups appear more prone to use research published in the national languages compared to English. These findings support recent claims for promoting multilingualism in scholarly communication (www.helsinki-initiative.org). Both OA and language diversity are needed to support the broader societal impact agendas of responsible research and innovation and open science.

The approach of using an online questionnaire, which targeted active users of the Journal.fi platform, was an effective method. However, the self-selection of participants may lead to the underrepresentation of certain groups. More detailed studies could look at the wider use of research from different academic disciplines. This study focused on the relatively small context of Finland, and similar approaches for studying other countries and regions would be beneficial for understanding the role of academic publishing in national languages more generally.

References

Cameron-Pesant, S. (2019). Usage et diffusion des revues savantes québécoises en sciences sociales et humaines : analyse des téléchargements de la plateforme Érudit. *Recherches sociographiques*, 59, 365-384.

Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. *Scientometrics*, 101, 1419-1430.

Haustein, S. (2014). Readership metrics. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond bibliometrics: harnessing multidimensional indicators of performance* (pp. 327–344). Cambridge, MA: MIT Press.

Kulczycki, E., Guns, R., Pölönen J., Engels, T., Rozkosz, E., … Sivertsen, G. (2020). Multilingual publishing in the social sciences and humanities: A seven-country European study. *Journal of the Association for Information Science and Technology*, 71, 1-15.

Kurtz, M. J. & Bollen, J. (2010). Usage bibliometrics. *Annual review of information science and technology*, 44(1), 1-64.

Late, E., Korkeamäki, L., Pölönen, J. & Syrjämäki, S. (2020). The role of learned societies in national scholarly publishing. *Learned Publishing*, 33, 5-13.

Mohammadi, E., Thelwall, M., Haustein, S. & Larivière, V. (2015). Who Reads Research Articles? An Altmetrics Analysis of Mendeley User Categories. *Journal of the Association for Information Science and Technology*, 66, 1832-1846.

Nicholas, D., Huntington, P, Dobrowolski, T., Rowlands, I., Jamali H. M. & Polydoratou, P (2005). Revisiting obsolescence and journal article decay through usage data: an analysis of digital journal use by year of publication. *Information Processing and Management*, 41, 1441-1461.

Stojanovski, J., Petrak, J. & Macan, B. (2009). The Croatian national open access journal platform. *Learned publishing*, 22, 263-273.

Wouters, P. & Costas, R. (2012). *Users, narcissism and control – Tracking the impact of scholarly publications in the 21st century*. Utrecht: SURF-foundation.

Zuccala, A. (2009). The lay person and Open Access. *Annual Review of Information Science and Technology*, 43(1), 1.

Exploring the Distribution Regularities of Referees' Comments in IMRaD Structure of Academic Articles

Chenglei Qin[1] and Chengzhi Zhang[2]

{[1] clqin, [2] zhangcz}@njust.edu.cn
Department of Information Management, Nanjing University of Science & Technology, Nanjing 210094 (China)

Introduction

Peer review is playing an essential role in scientific communication, which is closest to the natural state of the evaluated object (Narin, 1978). The peer review comments, as a carrier to record the process of peer review, objectively record the referees' suggestions, expert knowledge, and other valuable information. Most scientific articles' structural functions can be divided into IMRaD (Introduction, Materials & Methods, Results, and Discussion) structures. Exploring the distribution regularities of peer review comments among different structural functions can reveal the critical points of referees' focus and help students or early-career researchers deepen their understanding of peer review mechanisms. For a long time, limited by the traditional peer review mechanism, the peer review reports are not large-scale open access, scholars cannot unveil the mystery of peer review from the perspective of text content.

To further improve the fairness and transparency of the peer review process, improve the scientific quality, and speed up scientific dissemination, the open peer review mechanism came into being (Pöschl, 2012). Several influential journals, well-known publishing groups, and scientific research institutions began to adopt the open peer review mechanism (ASApbio, 2018; Madison Crystal, 2019). The open-access of many of peer review reports provides a data basis for mining and analyzing the peer review comments.

Based on the papers and peer review reports published in Atmospheric Chemistry and Physics (ACP, IF 5.6) from 2001 to 2016, this paper makes an exploration of the distribution of peer review comments in different structural functions of academic text to reveal the critical points of referees focus and provide reference for the writing of academic papers.

Methodology

Dataset

1,333 papers' section structural functions were recognized by the feature words of section type (Note that the feature words of section type can be found in Table 1 of Appendix, https://github.com/kakabular/peer-review/blob/main/Appendix_Table_1.pdf). The distribution of the dataset from 2001 to 2016 is shown in Figure 1.

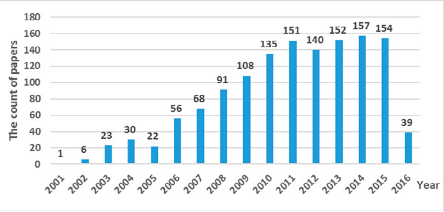


Figure 1. The distribution of papers from 2001 to 2016

The location information (such as page, line, table, figure, equation, etc.) of review comments can be extracted through the rules of the regular expression (The rules can be found in https://github.com/kakabular/peer-review).

According to the style of the ACP review reports, one paragraph generally comments on one problem. Therefore, we treat one paragraph as one comment sentence. The distribution of the number of peer review comments with location information is shown in Figure 2. The lowest coverage rate is 0.77.

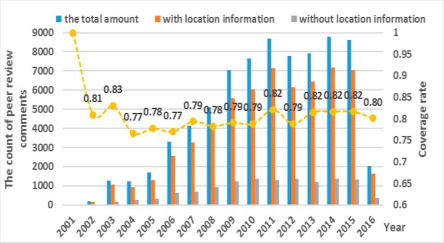


Figure 2. The distribution of peer review comments from 2001 to 2016.

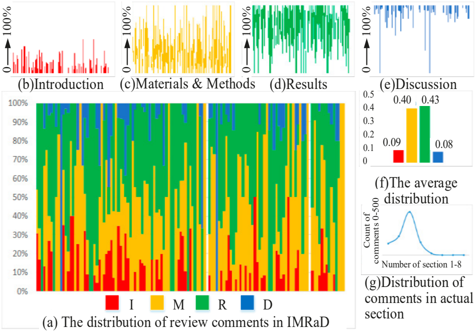
Method

This paper's research plan is as follows: First, we obtained research data from the ACP journal website. Because the original format of the manuscripts and the peer review reports are in PDF format, the PDF format files were converted into HTML files, and then the plain text is automatically extracted according to HTML source code. Second, use feature words of section type to recognize the academic articles' section structural function. And then, extract the location information in peer review comments according to the rules, maps the peer review comments to the original paper, obtains the corresponding structural function of the peer review comment. Finally, we analyze the peer review comments' distribution in academic articles' different structural functions (IMRaD) and actual sections.

Results

Taking the year of 2007 as an example, this section first analyses the distribution of peer review comments in the different structural functions of academic articles. Secondly, we analyse the average distribution of review comments in different years and the distribution of review comments in the actual sections.

Figure 3(a)- 3(e) shows the distribution of peer review comments in the IMRaD structure in 2007, which presents that in the majority of cases the proportion of peer review comments distributed in the Materials and Methods sections is significantly higher than in the Introductions and Discussion sections, reflecting deeper concern of the referees on materials, methods and experimental results (The abscissa represents each paper in 2007. The vertical axis represents the distribution percentage of review comments for each paper in IMRAD). Figure 3(f) shows the average distribution of the review opinions in the IMRaD structure that 40% and 43% of reviewer comments are directed at Materials & Methods and Results, which can verify the above findings. Figure 3(g) shows the distribution of the review comments in the actual sections, from which it can be seen that the number of review comments distributed in the opening and ending sections of the paper is significantly less than the sections in the middle part. Generally, scientific papers are organized according to the logical structure of IMRaD. So that we can get similar results.



(b)Introduction (c)Materials & Methods (d)Results (e)Discussion

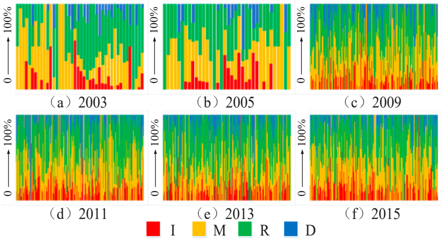
(a) The distribution of review comments in IMRaD

(f)The average distribution

(g)Distribution of comments in actual section

Figure 3. The distribution of peer review comments of data in 2007.

The distributions of peer review comments in other years (for example, in the year of 2003, 2005, 2009, 2011, 2013, 2015) are shown in Figure 4. We can find that referees pay more attention to the function of Materials & Methods and Results.



（a）2003 （b）2005 （c）2009

（d）2011 （e）2013 （f）2015

Figure 4. The distribution of peer review comments in other years

Conclusion

Based on ACP data from 2001 to 2016, this paper explores the distribution of peer review comments in different structural functions of academic articles and actual chapters. The results show that the reviewers pay more attention to Materials & Methods and Results. The proportion of reviewers' opinions in the middle sections of a paper is higher than that in the beginning and end of a paper. The findings are consistent with the general cognition: the scientific, reliable, and enough experimental methods are the premise of the research conclusion. This finding helps researchers understand the key points that referees focus on in reviewing processes, deepen their understanding of the review mechanism, and improve their writing skills. In scientific research, researchers should explore boldly and verify carefully to ensure the quality of science. Also, there are many shortcomings in this paper. For example, due to the significant differences in the writing style of the review experts, the rules cannot cover all the location information.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 72074113).

References

ASApbio. (2018). Open Letter on The Publication of Peer Review Reports. Retrieved December 18, 2020 from: Https://Asapbio.Org/Letter.

Madison Crystal. (2019). PLOS Journals Now OPEN for Published Peer Review. Retrieved December 18, 2020 from: Https://Theplosblog.Plos.Org/2019/05/Plos-Journals-Now-Open-For-Published-Peer-Review/

Narin, F. (1978). Objectivity Versus Relevance in Studies of Scientific Advance. *Scientometrics*, 1(1), 35-41.

Pöschl, U. (2012). Multi-Stage Open Peer Review: Scientific Evaluation Integrating the Strengths of Traditional Peer Review with The Virtues of Transparency and Self-Regulation. *Frontiers in Computational Neuroscience*, 6, 1-16.

References to Literature in Patent Documents: A Case Study of Electromobility Research

Zhao Qu

zhaoqu@dzhw.eu
Department 2, German Centre for Higher Education Research and Science Studies (DZHW), Schützenstr. 6a,
Berlin 10117 (Germany)

Introduction

Electromobility (e-mobility) takes a systematic standpoint that reaches far beyond mere technical aspects and also requires complex social changes (Grauers et al., 2013). Specifically, it is associated with the shift to a broader, more diverse and global network of actors and stakeholders, which may help promote and introduce innovations to the market (Capgemini, 2012). The crucial point for fostering innovation is the transfer and recombination of knowledge from numerous actors, thereby blurring the boundaries between science and technology (Campbell et al., 2005).

This study is to explore the connections between science and technology through journal articles pertaining to e-mobility cited in patent literature. This special issue is a bridging effort to bring together scientific papers, patent citations and comparative analysis by shining a spotlight on measuring the progress in knowledge generation, diffusion and transfer

Data and methods

This study consists of two stages of analysis, including an overview on global trends in e-mobility research as well as a further discussion on partial articles that are cited as non-patent literature (NPL) references and the matched citing patents. We limit our analysis to 3,441 journal articles and 1,162 patent documents retrieved from Web of Science and Lens patent corpus (Figure1).

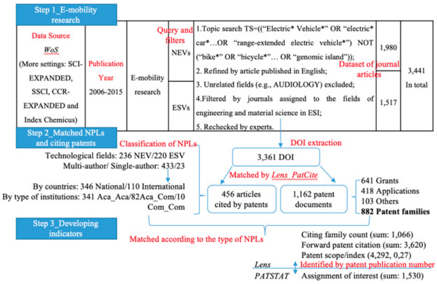

Step 1_E-mobility research
Data Source
WoS
(More settings: SCI-EXPANDED, SSCI, CCR-EXPANDED and Index Chemicus)
Publication Year
2006-2015
E-mobility research
Query and filters
NEVs
ESVs
1.Topic search TS=((“Electric* Vehicle*” OR “electric* car*…OR “range-extended electric vehicle*”) NOT (“bike*” OR “bicycle*”… OR “genomic island”));
2. Refined by article published in English;
3. Unrelated fields (e.g., AUDIOLOGY) excluded;
4.Filtered by journals assigned to the fields of engineering and material science in ESI;
5. Rechecked by experts.
Dataset of journal articles
1,980
1,517
3,441
In total
Step 2_Matched NPLs and citing patents
Classification of NPLs
Technological fields: 236 NEV/220 ESV
Multi-author/ Single-author: 433/23
By countries: 346 National/110 International
By type of institutions: 341 Aca_Aca/82Aca_Com/10 Com_Com
DOI extraction
3,361 DOI
Matched by Lens_PatCite
456 articles cited by patents
1,162 patent documents
641 Grants
418 Applications
103 Others
882 Patent families
Step 3_Developing indicators
Matched according to the type of NPLs
Citing family count (sum: 1,066)
Forward patent citation (sum: 3,620)
Patent scope/index (4,292, 0,27)
Lens
Identified by patent publication number
PATSTAT
Assignment of interest (sum: 1,530)

Figure 1. Dataset of e-mobility research and matched citing patents.

Three widely used indicators based on the existing literature, including the patent scope (Lerner, 1994), patent family size (Harhoff et al., 2003) and forward citations (Trajtenberg, 1990) are employed to capture the technological importance of the invention citing scholarly works and the possible impact on subsequent technological developments. NPL references are further linked to data on applicants and frequently transferred patents for the purpose of tracing technology transfer with a knowledge base provided by different types of collaboration.

Results and discussion

Global e-mobility studies have presented a growing trend both in counts and scientific collaborations since 2006, increasing both national and international links. The United States and China are the main sources of e-mobility research among 75 countries and territories identified by postal addresses of listed authors. Authors from 69 countries have engaged in an international collaboration, and when comparing the structure of institutional collaboration network, over 92 percent of total institutions have collaborated with others.

There is a generally upward trend in the number of citations received three years after publication except 2008. E-mobility studies cited in patent literature cover a smaller proportion by comparison to articles that do not match to NPL references, while the percentage of collaborative works is higher than that of all relevant publications during the same period. NPL references present a distribution pattern over time similar to that of the whole set. That is, the more publications a year or a country corresponds to, the higher the number of articles cited by patents is. However, the rank of countries by the percentage of articles cited by patents displays an entirely different picture (Table1).

The publication date of citing patents ranges from 2008 to 2018, and the coverage of patented technologies is comparatively broad, ranging from human necessities to emerging cross-sectional technologies. Patents citing e-mobility research are principally applied by companies in the United States, Singapore and Japan. Seventy-six percent of patents have cited scientific papers with a collaboration between academic institutions corresponding to 73 percent of applicants from companies and academia (Figure 2). Articles containing an academia-industry collaboration have been largely cited in company’s patent literature while the one published by companies is more likely to be cited by academic patents. Patents citing research conducted through an academia-industry collaboration have more forward citations, while patents with citations to the collaboration between authors from companies cover a broader scope of technologies (Table 2). More than half of citing patents have been assigned to new parties, and 75% of assignments are made from individuals to companies or universities. Eighty percent of patents assigned more than ten times have cited collaborative articles of academic institutions.

Table 1. Distribution of e-mobility research cited by the patent.

Country by publication count	Country by NPL count	Country by NPL percentage
US/872	US/160	BG/40.00%
CN/704	CN/79	CL/33.33%
KR/248	KR/38	FJ/33.33%
UK/207	DE/30	PY/33.33%
FR/206	GB/28	NZ/28.13%
JP/191	FR/26	IL/25%
DE/280	JP/25	RU/25%
IT/162	CA/19	TH/20%
CA/158	IT/18	FI/17.39%
ES/125	ES/16	CH/16.39%

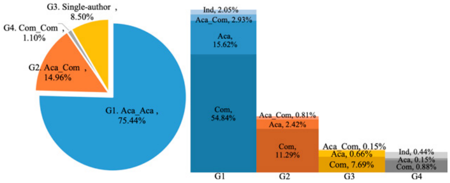


Figure 2. Patents with different NPL citations and matched groups of applicants.

Table 2. Indicators based on citing patent with different NPL citations (Average, SE).

NPL	*Patent family size*	*Patent scope index*	*Forward citation*	*Patent assignment*
NEV	7.46 (0.35)	0.25 (0.01)	2.08 (0.37)	2.09 (0.10)
ESV	7.65 (0.25)	0.29 (0.01)	3.25 (0.25)	2.47 (0.25)
Sin-Au	7.57 (0.51)	0.32 (0.02)	3.84 (1.12)	2.87 (0.80)
Mul-Au	7.43 (0.23)	0.27 (0.01)	2.76 (0.24)	2.52 (0.11)
National	7.62 (0.26)	0.26 (0.01)	2.43 (0.29)	2.05 (0.08)
International	7.74 (0.45)	0.27 (0.02)	3.09 (0.31)	2.98 (0.59)
Aca_Aca	7.24 (0.27)	0.26 (0.01)	2.54 (0.29)	2.24 (0.12)
Aca_Com	6.74 (0.30)	0.26 (0.02)	2.65 (0.42)	1.97 (0.22)
Com_Com	7.07 (1.75)	0.27 (0.07)	1.40 (0.22)	1.57 (0.36)

Conclusions

This study has provided an informative overview of e-mobility research and matched citing patent documents. Global research on e-mobility since 2006 has been conducted increasingly by a broader network of authors. In the light of evidence from patent citations, e-mobility research cited frequently by other articles are more likely to be cited by patents. Articles exploring the issues of EVs published by universities or research institutes, especially with the national collaboration, have received more patent citations. Mostly, academic institutions' collaborative articles have been cited by applicants from companies and academia. Studies conducted by an academia-industry collaboration have been widely cited in company's patents while the one with a business collaboration is more likely to be cited by academic patents. Patents citing single-author articles and international research collaborations seem to be more important indicated by broader patent scope, larger family size and more forward citations on average. More than half of citing patents have been assigned primarily from individuals to companies or universities, and most of frequently transferred patents have cited collaborative articles of academic institutions. Differences mentioned above need to be explored in a broader context, while the growing trend in scientific collaborations of NPL references and its influences on patents measured in each dimension should not be neglected.

Acknowledgments

The author would like to thank express gratitude for the financial supports provided the National Science Foundation of China (No.71673036) and the Chinese Academy of Engineering (2016-XZ-03-05).

References

Campbell, D. F. & Guttel, W. H. (2005). Knowledge production of firms: research networks and the" scientification" of business R&D. International *Journal of Technology Management*, 31(1-2), 152-175.

Capgemini. (2012). *Managing the Change to E-Mobility*. Retrieved December 20, 2017 from: https://www.capgemini.com/wp-content/uploads/2017/07/Managing_the_Change_to_eMobility_Capgemini_Automotive_Study_2012.pdf

Grauers, A., Sarasini, S. & Karlström, M.(2013). Why electromobility and what it is? In Sandén, B. (Ed.), *Systems Perspectives on Electromobility* (pp. 10-21). Chalmers University of Technology, Göteborg.

Harhoff, D., Scherer, F. M. & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy*, 32(8), 1343-1363.

Lerner, J. (1994). The importance of patent scope: An empirical analysis. *RAND Journal of Economics*, 25(2), 319-333.

Trajtenberg, M. (1990). A Penny for Your Quotes: Patent Citations and the Value of Innovation. *RAND Journal of Economics*, 21(1), 172-187.

The scientific productivity of German PhD graduates: A machine learning–based author name disambiguation and record linkage approach

Andreas Rehs[1]

[2] *andreasrehs@googlemail.com*
University of Kassel, International Center for Higher Education Research, Mönchebergstr. 17, 34109 Kassel (Germany)

Introduction

Although PhD students and graduates play an important role in scientific knowledge production and economic growth (Stephan, Sumell, Black, & Adams, 2004), we have little information on their publications and related bibliometric indicators. Therefore, there is a need for new data sources to identify PhD students' publications and a need to link them to other databases of interest. This especially holds for Germany where the problem of missing publication information on PhD graduates has repeatedly been pointed out (Konsortium Bundesbericht Wissenschaftlicher Nachwuchs, 2017, p. 35).

Data

The dataset we used to trace the publication record of German PhD graduates is built on two sources: the electronic catalog of the German National Library (DNB) and the Web of Science (WOS). The DNB is mandated to collect all German dissertations, and therefore its catalog lists the vast majority of PhD theses submitted at German universities. For every dissertation, the DNB stores some basic information, like the dissertation's year, the granting university and the author's name. Regarding our second dataset, we use a 2017 copy of the WOS as the basis to retrieve publication data.

Author name disambiguation

To resolve ambiguous author names in the WOS, we perform author name disambiguation on the basis of Rehs (2021). In this task, we consider all surname-initial combinations (in the following: blocks) that appear in the DNB dissertation database as worthwhile to disambiguate. There are 533,198 distinct blocks in the DNB; 153,213 appear more than once, which means there is more than one dissertation related to that block. In a 30-day processing period, the disambiguation algorithm processed 184,783 of the 533,198 blocks in the DNB and ended up with 10.6 million publications processed into about 1.96 million different authors. These processed homonyms cover about 51% of the 960,000 relevant dissertations in the DNB. We consider dissertations between 1975 and 2015 and those outside the disciplines: history, philosophy, theology, and arts and music as relevant. Concerning the complete WOS, our disambiguation approach processed 2.8% of the 6.5 million homonyms. Those homonyms account for 4.2% of the 178 million block-publication relationships in the WOS (Rehs, 2021).

Probabilistic record linkage of WOS data to German dissertation authors

In the next step, we want to connect entries in the DNB with those in our disambiguated WOS database. However, we lack unique identifiers in the two databases and need to apply record linkage. Record linkage is either deterministic or probabilistic. In deterministic linkage, records are matched if linkage fields agree, or unmatched if they disagree. Deterministic approaches are not robust to measurement error (e.g., related to misspellings) and missing data. Additionally, uncertainty in the merging procedure cannot be quantified. Instead, arbitrary thresholds determine the similarity sufficient for matches.

Probabilistic record linkage approaches can account for this uncertainty (Fellegi & Sunter, 1969). They estimate the probability that two given records refer to the same or different entities. In this process, each identifier, such as first names, is weighted by its performance in determining matches and non-matches between two databases. For the determination of weights, two probabilities are of interest: unmatched probability (u-probability) and matched probability (m-probability). The u-probability gives the probability that a variable between two datasets agrees by chance. For example, we use the address string to compare records between the DNB and the disambiguated WOS dataset that share the same homonym "Muller, M". If we assume ten unique and uniformly distributed addresses in the two databases, the u-probability would be 1/10, or 0.1. The m-probability describes the probability that a variable in matching pairs will agree. Again using the example of addresses and abstracting from peculiarities of the two datasets, the matched probability between two records may be exactly 1.

That means the address of a dissertation always agrees with the address stated on the corresponding publication. However, we don't know much about this probability and the related true matches. We can only estimate the m-probability using iterative methods, such as the expectation maximization algorithm (Dempster, Laird, & Rubin, 1977). Given that we would estimate 0.95 for the m-probability, the address variable weights would be calculated as described in Table 3. The described calculation can be repeated for all other possible variables between the two datasets. Finally, the total weight for matching and non-matching are determined by adding all variable weights. The posterior probability for matching is then derived from the total weight (Blakely & Salmond, 2002).

Table 1. Identifier weighting in probabilistic record linkage

Variable	**Outcome**	**Proportion of links**	**Proportion of non-links**	**Frequency ratio**	**Weight**
Address	***Match***	m= 0.95	u≈ 0.1	m/u≈ 9.5	ln(m/u)/ln(2) ≈ 3.25
Address	***Non-match***	1−m= 0.05	1-u≈ 0.9	(1-m)/(1-u) ≈ 0.05	ln((1-m)/(1-u))/ln(2) ≈ -4.17
First name	…				

Depiction based on (Blakely & Salmond, 2002).

In our approach, we use an extension of this framework (Enamorado & Fifield, 2019). This "fastlink" framework allows the incorporation of partial agreements and missing data, which relaxes the conditional independence condition of the Fellegi-Sunter model. The conditional independence condition requires the variables to be independent of each other, which is seldom the case with real-world data. In our case, a publication's address field and the author's first name, for instance, correlate in their submissions.

In the next step, we prepare the datasets and take into account that both datasets have peculiarities. While the DNB, as a dissertation database, usually captures a single event, our disambiguated WOS database covers a researcher's lifetime record of publications. Therefore, we aggregate all WOS publications belonging to a disambiguated author and build a synthetic author profile consisting of the minimum paper year, the mode value of the address, and the mode value of the first name.

In our further procedure, we consider only blocks with sufficient frequency. A minimum frequency is computationally needed to conduct reasonable probabilistic inference (Enamorado et al., 2020). We do so by setting the threshold to blocks with at least ten different dissertations and at least ten different disambiguated authors. In the last step we apply the fastlink framework to the the synthetic WOS author profiles and the DNB. We use the DNB university city vs. the mode value of the synthetic author profiles address, first name vs. first name and dissertation year vs. minimum value of paper year to link both datasets.

We apply deterministic linkage for blocks containing fewer observations than the mentioned thresholds. Here, we merge authors that match in the same variables. If there are fewer than three entries in both databases, we relax the matching to comparison of years and one other variable. Generally, we allow a margin of +/− 2 for year. For address and first name we allow fuzzy matching but allow variation of only one character in the Levinstein distance used.

Results

In our probabilistic approach, we were able to link 30,840 dissertation authors to corresponding author profiles. This number is based on all observations with higher than 0.8 posterior probability. The deterministic strategy yields 30,804 dissertations. Our dataset of 61,644 linked authors covers 5.4% of the relevant DNB dissertations and 0.03% of the disambiguated WOS author profiles.

References

Blakely, T., & Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predictive value. In Great Britain International Journal of Epidemiology (Vol. 31).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

Enamorado, T., & Fifield, B. (2019). Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. American Political Science Review, 113, 353–371.

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64(328), 1183–1210.

Konsortium Bundesbericht Wissenschaftlicher Nachwuchs. (2017). *Bundesbericht Wissenschaftlicher Nachwuchs 2017*. Bielefeld: W. Bertelsmann.

Rehs, A. (2021). A supervised machine learning approach to author disambiguation in the Web of Science. *Journal of Informetrics* (forthcoming).

Stephan, P. E., Sumell, A. J., Black, G. C., & Adams, J. D. (2004). Doctoral education and economic development: The flow of new Ph.D.s to industry. Economic Development Quarterly, 18(2), 151–167.

Do COVID-related articles pass through the fast lane? The case of the International Journal of Infectious Diseases

Ronald Rousseau

ronald.rousseau@uantwerpen.be
Faculty of Social Sciences, University of Antwerp, Middelheimlaan 2, 2020 Antwerpen (Belgium)

ronald.rousseau@kuleuven.be
Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, Naamsestraat 61, 3000 Leuven (Belgium)

Introduction

It has been stated that in 2020 many journals, especially biomedical and multidisciplinary ones, have accelerated their throughput, especially for COVID-19 related submissions (Callaway, 2020; Horbach; 2020). Horbach (2020) analysed the duration of the publication process, in number of days, for a sample of 529 journal articles, published in 14 different journals, including the *International Journal of Infectious Diseases* (IJID). For this investigation, he used a repository of coronavirus-related research articles established by the Centre for Science and Technology Studies (CWTS), more specifically the April 4, 2020 version. The main conclusions of his study are that medical journals have, since the outbreak of the pandemic, strongly accelerated their publication process for coronavirus-related articles. Concretely, he found that the time between submission and publication has decreased on average by 49%. The largest decrease in number of days between submission and publication of articles was due to a decrease in time required for peer review. For articles not related to COVID-19, no acceleration of the publication process was found.

For IJID he considered 32 articles published before October 1, 2019, and 32 dealing with COVID-19 published after January 1 (and before April) 2020. As a comparison, he further considered the ten most recently published articles (as of April 16, 2020) dealing with studies not related to COVID-19.

In this submission, we focus on IJID. This journal, the flagship journal of the International Society for Infectious Diseases, is a truly international journal to which many non-Western and non-Chinese authors contribute. Concretely, we only consider publications of article type (according to the WoS classification) and made a distinction between COVID-related articles and other ones. In 2020, 709 documents of article-type were published in this journal.

Research questions

1) Is there a noticeable increase in the speed with which articles are published in IJID?

2) If there is an increase, does it apply to all articles, or only to COVID-related ones?

3) If there is an increase is it due to the time between submission and acceptance, or the time between acceptance and online availability, or both?

Methods

COVID-related articles were determined based on the WoS query:

TS= (“SARS-CoV-2” OR COVID* OR “2019-nCoV”)

Other articles were considered not COVID-related. As we collected data manually we were able to check that this query indeed retrieved all COVID-related articles (and we again noticed problems with the WoS definition of reviews (Colebunders & Rousseau, 2015), see below). As the baseline, we consider all articles in volumes 90 and 91 (published in January and February 2020). Then we collected all COVID- and non-COVID articles submitted between January 23, 2020 (first submission of a COVID-related article in IJID) and June 30, 2020. For each of these articles, we noted the submission date, the acceptance date, and the date the articles became available online. The time in days between any two dates was determined (taking into account that 2020 was a leap year). The most recent articles, among those submitted before July 2020, were published in the February 2021 volume.

Results

Results for the time between submission and acceptance, here referred to as acceptance delay, are shown in Table 1 while the corresponding data for time between acceptance and online availability (availability delay) are shown in Table 2.

Although these data clearly show that the average acceptance delay is much smaller for COVID-related articles than for non-COVID-related ones, we see that also for non-COVID-related ones the delay was shortened compared to the baseline period. Availability delay is always very short for this journal and remained practically the same, yet slightly shorter for COVID-related articles than for the others. In this sample of IJID articles, the decrease in average acceptance delay for COVID-19 articles is 43% while for non-COVID-related articles it is 9%.

Table 1. Acceptance delay, time in days

	# art.	Mean delay	St.dev	Md.	Q1
Baseline	103	71.02	38.05	68	[3,34]
COVID	214	40.35	32.65	31	[1,19]
Non-COVID	161	64.92	39.06	55	[1,36]

Note: Md. stands for median and Q1 for first quarter

Table 2. Availability delay, time in days

	# art.	Mean delay	St.dev	Md.
Baseline	103	6.92	5.40	6
COVID	214	7.38	7.52	5
Non-COVID	161	7.55	8.37	6

Applying a Mann-Whitney test, as described e.g. in (Rousseau et al., 2018, p. 95) the null hypothesis of equality of distributions of acceptance delays was rejected for two cases: baseline versus COVID-related, and COVID versus non-COVID articles (z-scores larger than 7 in both cases). The z-score for the baseline versus the non-COVID case is 1.636. Hence we cannot reject the null hypothesis that these two distributions are the same. Yet, there is no reason to think that speeding up has occurred at the cost of non-COVID-related papers.

Discussion

Although we come essentially to the same conclusion as Horbach (2020) for IJID, our findings are based on a much larger dataset, covering a longer period.

The acceptance period involves three types of actors: the reviewers, the authors revising their manuscript, and the editor(s) handling the manuscript. For COVID-related papers, all three groups have a motif to speed up their actions and we can only assume that they actually did.

The reader may see in Table 1 that there is one non-COVID-related article that was accepted after one day. This article is published as a review, but considered an article by Clarivate Analytics. It is a discussion of a practical toolkit for developing Antimicrobial Stewardship Programs (ASPs) developed by the WHO (Pierce et al., 2020). Probably the main editor immediately accepted it.

We offer the following thoughts about the differences in time between the three groups of papers. Assume that COVID-related papers are treated with preference compared to non-COVID-related ones. Then borderline COVID papers may yet be accepted (because of their possible scientific interest) but need a longer time to be revised. This would lead to a higher average acceptance time than would otherwise be the case. Assume further that for non-COVID-related papers the opposite is the case, so that borderline papers are rejected. This would shorten the average acceptance time and hence brings the two acceptance times closer to one another. Regrettably, we have no information on rejection rates, which is another important parameter when studying journals and their publications. The first quarters, though, (Table 1) show that there are much more COVID papers, compared with the other two groups, that are accepted in a very short time.

Conclusion

IJID indeed accepted COVID-related submissions in the first half of 2020 faster than it did earlier papers. There is no indication that this happened at the cost of other submissions. On the contrary, also other articles were published faster, although not in a statistically significant way.

Speeding up the time between submission and online publication was mainly due to a decrease in acceptance times as this journal has already a very short availability time.

Acknowledgments

The author thanks Robert Colebunders (University of Antwerp) for suggesting this investigation.

References

Callaway, E. (2020). The COVID-19 crisis could permanently change scientific publishing. *Nature*, 582(7811), 67-68.

Colebunders, R. & Rousseau, R. (2013). On the definition of a review, and does it matter? *Proceedings of ISSI 2013 Vienna,* (Juan Gorraiz, Edgar Schiebel, Christian Gumpenberger, Marianne Hörlesberger, Henk Moed, eds.), AIT Austrian Institute of Technology, Vienna, p. 2072-2074.

Horbach, S.P.J. M. (2020). Pandemic publishing: medical journals strongly speed up their publication process for COVID-19. *Quantitative Science Studies*, 1(3), 1056-1067.

Pierce, P., Apisarnthanarak, A., Schellack, N., Cornistein, W., Al Maanie, A., Adnan, S. & Stevens, M.P. (2020). Global antimicrobial stewardship with a focus on low- and middle-income countries: A position statement for the international society for infectious diseases. *International Journal of Infectious Diseases*, 96, 621-629.

Rousseau, R., Egghe, L. & Guns, R. (2018). *Becoming metric-wise. A bibliometric guide for researchers*. Kidlington: Chandos (Elsevier)

Editorial recommendations in Nature and Science

Jasmin SADAT[1] and João MARTINS[2]

[1] *jasmin.sadat@ec.europa.eu*
European Research Council Executive Agency, A1 – Support to the Scientific Council,
1049 Brussels (Belgium)

[2] *joao.martins@snf.ch*
Swiss National Science Foundation, Wildhainweg 3, 3001 Berne (Switzerland)

Introduction

Due to the steady increase of scientific literature, one important question that arises is how to identify outstanding contributions in a continuously growing pool of scientific publications?
Over the years, several channels dedicated to signaling outstanding publications in the immense collection of research knowledge have evolved in the scientific community (Mugabushaka, Sadat, & Costa Dantas Faria, 2020). Up to date, the most common way to identify outstanding publications relies on bibliometric indicators, even though there is criticism to this approach (e.g., Aksnes, Langfeldt, & Wouters, 2019). Here we will focus on an alternative way to identify such publications, namely by exploring scientific expert judgements of two major science journals: *Nature* and *Science*.

In recent years, a number of scientific journals have added sections dedicated to provide better visibility for outstanding research from among the high number of publications. For example *Nature* and *Science* provide a weekly selection for especially outstanding contributions from among all of their own journals' publications (e.g., *Nature*'s "News & Views" or *Science*'s "This Week in Science") and from the pool outside of their own journal (e.g., *Nature*'s "Research Highlights" or *Science*'s "Editor's Choice"). These highlighted sections, so-called editorial recommendations, are usually curated by scientific editors and refer to scholarly articles that are judged as being highly influential and promising for future research developments. They carry a precious expert judgement of scientific quality and are readily available unlike current bibliometric evaluations. Thus, such featured publications provide valuable information for research monitoring.
Up to now, this type of scientific recognition has rarely been used or described in detail. In what follows, we will outline some main characteristics for two such example datasets collected from *Nature* and *Science*. We will focus on the journals of publication to which the editorial recommendations refer to, the main authorship countries, and a tentative exploration with respect to their research areas. This will increase our understanding of similarities or differences dependent on the highlighting journal and further describe important characteristics for this type of scientific recognition.

Methods

In order to compare editorial recommendations of different journals, they need to have the same regularity, scope, and target audience (Sadat & Mugabushaka, 2020). Thus, we focused on weekly occurring, outward directed editorial recommendations with a similar target audience: (a) *Nature*'s "Research Highlights" and (b) *Science*'s "Editor's Choice". These two weekly sections each highlight around seven publications per issue, coming from across a wide range of research areas, pointing to scientific journals outside their own, and are targeted to the scientific community. Datasets of editorial recommendations were collected from the information available on the journals' websites. The collection of Digital Object Identifiers (DOIs) for scholarly articles was in most cases automated. However, automated retrieval was always followed by a substantial effort in manual data cleaning and curation, since many references to DOIs were missing or unstructured.
For (a), we collected and cleaned a dataset with 5'345 editorial recommendations of scholarly publications for a period from January 2008 to November 2020. For (b), we collected and cleaned a dataset with 4'085 editorial recommendations of scholarly publications for a period from January 2009 to May 2020 (data available here: http://doi.org/10.5281/zenodo.4707580). To obtain additional information about the referenced publications, we retrieved meta data from CrossRef for each DOI regarding the journal (5'323 for [a] and 3'999 for [b]) and the country of authors from Scopus (4'302 for [a] and 2'981 for [b]). To gain a better picture of the dominant research areas in editorial recommendations, we present some preliminary analysis regarding the field information data provided by the journal (5'294 DOIs across 517 different fields for [a] and 4'083 DOIs across 875 fields for [b]). Due to their diversity, they were regrouped into the following most frequent categories based on subjective judgement: Astronomy, Biology, Chemistry, Ecology, Genetics, Geosciences, Materials, Medical research,

Microbiology, Neuroscience, Physics, Social Sciences, Technology, and Zoology.

Results

Which journals are most commonly referenced?

The main journals of editorial recommendations from *Nature* and *Science* are typical high impact journals (e.g., PNAS and Cell; see Figure 1). The highest share of *Nature*'s "Research Highlights" refers to publications in *Science*, whereas *Science*'s "Editor's Choice" fails to show the reverse pattern.

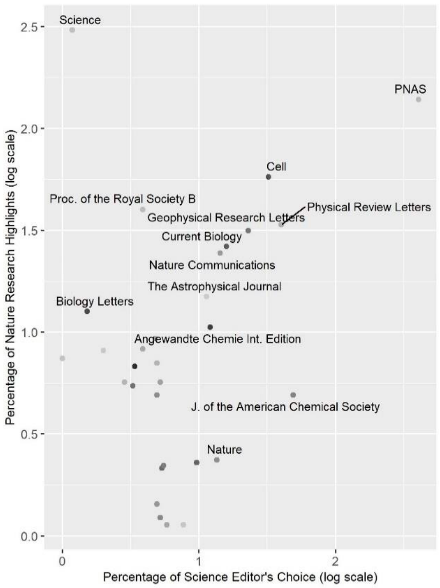


Figure 1. Percentage of referenced journals in editorial recommendations.

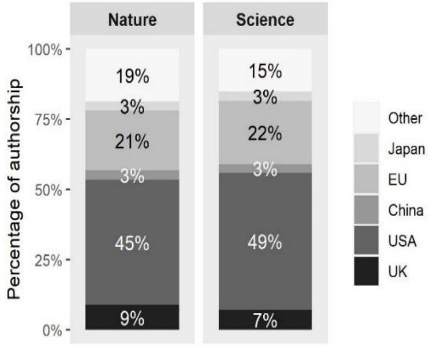


Figure 2. Percentage of authorship countries in editorial recommendations.

Which authorship countries are most common?

Geographical distribution of authorship was calculated as fractional percentage on a whole-normalized country-based level and was similarly distributed in both journals (see Figure 2). USA-based scientists author the main share of referenced publications in the editorial recommendations of *Nature* and *Science*, preceding EU-based ones and a group of scientists from countries other than the listed ones.

Which research areas are most common?

The distribution of editorial recommendations across research areas was similar in *Nature* and *Science*, apart from a few exceptions (see Table 1).

Table 1. Diverging research areas for editorial recommendations

Journal	Biology	Chemistry	Ecology	Zoology
Nature	9.5 %	4.5 %	14.2 %	6.7 %
Science	17.1 %	7.5 %	10.3 %	0.7 %

Outlook

The current poster aims at describing the main characteristics of editorial recommendations from two major science journals. We showed similarities and differences regarding the referenced publications, their authorship countries, and their main areas of research. Next steps of analysis entail the systematic comparison of editorial recommendations with concurrent measures of research impact like citation metrics and internet-based mentions. In sum, the current project provides a first description of outstanding publications highlighted in two major science journals, providing a promising new approach to evaluating research output.

Acknowledgments

We are grateful to Dr. Alexis-Michel MUGABUSHAKA for helpful comments and discussions about this work.

References

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. SAGE Open. https://doi.org/10.1177/2158244019829575

Mugabushaka, A.-M, Sadat, Costa Dantas Faria, J. (2020). In Search of Outstanding Research Advances: Prototyping the creation of an open dataset of "editorial highlights". arXiv:2011.07910

Sadat, J., & Mugabushaka, A.-M. (2020). In Search of Outstanding Research Advances – Exploring Editorial Recommendations. Zenodo. http://doi.org/10.5281/zenodo.4155204

Watching over innovation studies: Profiling the gatekeepers

Ana Teresa Santos[1] and Sandro Mendonça[2]

[1] *atmss@iscte-iul.pt*
Instituto Universitário de Lisboa (ISCTE-IUL) – Business Research Unit (BRU-IUL), Lisbon (Portugal)

[2] *sfm@iscte-iul.pt*
Instituto Universitário de Lisboa (ISCTE-IUL) – Business Research Unit (BRU-IUL), Lisbon (Portugal)

Introduction

Academic serials (especially peer-reviewed journals) play a very critical role in the scientific ecosystem and both integrity and independence are perceived as essential for good editorial governance (Rynes, 2006). For being responsible for articles selection (Bedeian et al., 2009; Feldman, 2008), elite board membership ensures the scientific quality of publications and "occupy key roles as opinion formers, gatekeepers and arbiters of disciplinary values" (Burgess and Shaw, 2010, p.630). So far, board elites have not been subject to a scrutiny proportional to their decision power (Burgess & Shaw, 2010) and an overall lack of transparency has been reported about the general editorial process despite of being actual gatekeepers (Miner, 2003; see also Bedeian et al., 2009; Horan et al., 1993).

In this work, we draw on the existing, but limited, body of knowledge available to examine empirically journal editorial boards (EB) in innovation studies field. We shed some light on demographics characteristics of editorial members. We believe this work may add some relevant information for those interested in science governance and social structures of research activities.

The Boards of "Innovation Studies"

Fagerberg et al. (2012) studied both the most productive journals in innovation as well as those using the most findings already published. For the 20 top-tier journals identified, we aim to present the demographic features of scholars behind. Scholars' names, affiliation country and gender were collected from official journal's editorial page. A total of 3,005 available seats were recorded occupied by 2,440 distinct persons from 30 countries.

The size of EBs

The mean size of Boards is 150 editors. Although *R D Manag* Board was found with only 19 editors, others have over 300 scholars: the likes of North-American outlets as *Acad Manage J* and *Acad Manage Rev* while *Manage Sci* stands out as the one with the largest team, 399 editors as presented in Figure 1.

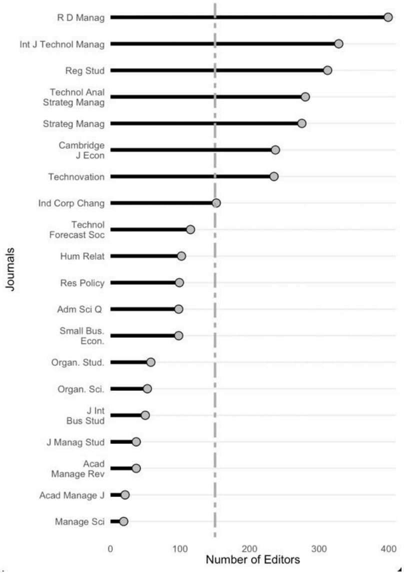


Figure 1: Editorial Board size per journal intercepted by the mean number of editors.

Geographies of editorship

Braun (2004) defined international journals as those with scholars from 5 countries in EB. Thus, all outlets from our study set are international ones, as their editors come from 8 to 30 different countries. Figure 2 shows a world representation of editors' frequency found in each country. Darker the colour, higher the frequency of editors.

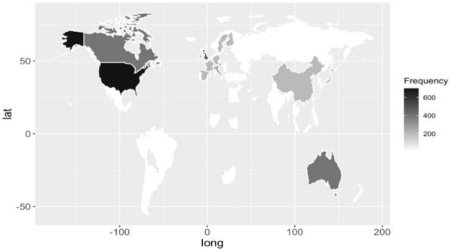


Figure 2. Geographical location of editorial members.

The US overwhelming dominance of membership is clear. García-Carpintero et al. (2010) reminded most science publishing houses are US-based. Europe and India host a significant number of editors. With exception of Australia, nations from the South represent a negligible role in this editorial process with very few editors involved.

Gender balance

With the purpose of understanding gender balance, we matched gender proportion in ten countries with higher numbers of editors as illustrated in Figure 3.

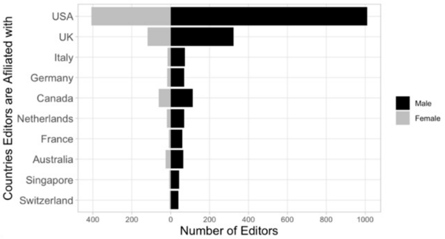


Figure 3: Number of editors affiliated to the top 10 countries, by gender.

US number of editors is by far different from the other countries with around 1,000 men and 400 women editors. However, US are not gender-balanced. Actually, all countries show a higher frequency of memberships held by men editors than women. Metz & Harzing (2009) pointed women's presence in academia is not long enough to reach levels of seniority which are associated with Board membership.

Conclusions

The present study is an attempt to understand the demographic structure of innovation EB members which showed to be diverse quantitative and qualitatively. Despite of being an eclectic topic, a low diversity for gender (male predominance) and country affiliation (high representation of US and UK editors) is perceived. A greater diversity is important for EB meet their missions (Jagsi et al., 2008) and a key driver in academia for knowledge development by applying different methodologies and paradigms (Robinson & Dechant, 1997).

References

Bedeian, A. G., van Fleet, D. D. & Hyman, H. H. (2009). Scientific Achievement and Editorial Board Membership. *Organizational Research Methods*, 12(2), 211-238. https://doi.org/10.1177/1094428107309312

Braun, T. (2004). Keeping the Gates of Science Journals. In *Handbook of Quantitative Science and Technology Research* (pp. 95-114). Springer Netherlands. https://doi.org/10.1007/1-4020-2755-9_5

Burgess, T. F. & Shaw, N. E. (2010). Editorial Board Membership of Management and Business Journals: A Social Network Analysis Study of the Financial Times 40. *British Journal of Management*, 21(3), 627-648. https://doi.org/10.1111/j.1467-8551.2010.00701.x

Fagerberg, J., Fosaas, M. & Sapprasert, K. (2012). Innovation: Exploring the knowledge base. *Research Policy*, 41(7), 1132-1153. https://doi.org/10.1016/J.RESPOL.2012.03.008

Feldman, D. C. (2008). Building and Maintaining a Strong Editorial Board and Cadre of Ad Hoc Reviewers. In *Opening the Black Box of Editorship* (pp. 68-74). Palgrave Macmillan UK. https://doi.org/10.1057/9780230582590_7

García-Carpintero, E., Granadino, B. & Plaza, L. (2010). The representation of nationalities on the editorial boards of international journals and the promotion of the scientific output of the same countries. *Scientometrics*, 84(3), 799-811. https://doi.org/10.1007/s11192-010-0199-3

Horan, J. J., Weber, W. L., Fitzsimmons, P., Maglio, C. J. & Hanish, C. (1993). Further Manifestations of the MOMM Phenomenon. *The Counseling Psychologist*, 21(2), 278-287. https://doi.org/10.1177/0011000093212011

Jagsi, R., Tarbell, N. J., Henault, L. E., Chang, Y. & Hylek, E. M. (2008). The representation of women on the editorial boards of major medical journals: A 35-year perspective. *Archives of Internal Medicine*, 168(5), 544-548. https://doi.org/10.1001/archinte.168.5.544

Metz, I. & Harzing, A.-W. (2009). Gender Diversity in Editorial Boards of Management Journals. *Academy of Management Learning & Education*, 8(4), 540-557. https://doi.org/10.5465/amle.8.4.zqr540

Miner, J. B. (2003). Commentary on Arthur Bedeian's "The Manuscript Review Process: The Proper Roles of Authors, Referees, and Editors." *Journal of Management Inquiry*, 12(4), 339-343. https://doi.org/10.1177/1056492603259056

Robinson, G. & Dechant, K. (1997). Building a Business Case for Diversity. *The Academy of Management Executive*, 11(3), 21-31. https://doi.org/10.2307/4165408

Rynes, S. L. (2006). "Getting on board" with AMJ: Balancing quality and innovation in the review process. *Academy of Management Journal*, 49(6), 1097-1102. https://doi.org/10.5465/AMJ.2006.23478050

Keyword Distance Ratio:
Evaluating Keyword Assignment with Word Embeddings

Brandon Sepulvado

sepulvado-brandon@norc.org
NORC at the University of Chicago, 4350 East West Highway, Bethesda, MD 20814 (USA)

Introduction

Despite the emergence of natural language processing (NLP) assisted search within certain new bibliometric databases (Wang et al., 2020), keywords remain a staple of information curation and retrieval within most major bibliometric sources, such as Web of Science, Scopus, and PubMed, and authors still choose keywords or phrases to summarize the information their documents contain. As such, keywords are an institutionalized component of the publication and literature search process, yet how does one empirically evaluate the keywords chosen for a document?

This text proposes the Keyword Distance Ratio (KDR) to respond to this question. Theoretically, keywords should, at the same time, be closely related to the content within a document while also remaining sufficiently distinct so as not to convey redundant information. The KDR is a document-level measure that relies upon the Relaxed Word Mover's Distance (RWMD; Kusner, Sun, Kolkin & Weinberger, 2015) to quantify these intuitions.

In investigating the utility of the KDR, this text compares article keywords submitted by authors to those chosen by Scopus in its index terms and then illustrates a dramatic difference in document summary based upon the keyword source.

Keyword Distance Ratio

Relaxed Word Mover's Distance

The KDR relies upon Word Mover's Distance (WMD; Kusner et al., 2015) in order to obtain distances between keywords and the documents they describe. For an intuitive understanding of WMD, imagine that a document is a cluster of points in n-dimensional space. The points represent the embeddings for words (i.e., keywords and those in the article), and a word's location is based upon the word's embedding vector. The current analyses use 300-dimension FastText embeddings (Mikolov, Chen, Corrado & Dean, 2013), which means that each word would be placed in a 300-dimension space. One plots the keywords and words in the document, and the distance then is a function of the effort it takes to move one set of points to the closest points in the other set. To reduce computational complexity, the RWMD relaxes a couple constraints imposed upon the WMD.

Keyword Distance Ratio

The KDR entails calculating two sets of RWMDs: (1) between each pair of keywords and (2) between each keyword and its corresponding document. The KDR then compares the sum of each type of distance, adjusting for the number of keywords listed. Let the KDR of document d be:

$$KDR_d = \frac{\sum_{j=1}^{n-1}\sum_{i>j} RWMD_{k_{i,d},k_{j,d}}}{\sum_{i=1}^{n} RWMD_{k_{i,d},a_d}} \cdot \frac{2}{(n-1)}$$

In this equation, the numerator sums all pairwise RWMDs between keywords k_i and k_j used to signify the content of scholarly document d; the denominator sums the pairwise RWMDs between keyword k_i and abstract a for the document d. This value however will increase automatically with each additional keyword because there are $\frac{n(n-1)}{2}$ possible comparisons in the numerator and only n comparisons in the denominator, where n is the number of keywords. As such, the second part of the KDR equation accounts for the number of keywords. To rephrase more intuitively, the KDR is the ratio of the sum of pairwise RWMDs for all keywords listed for a document to the sum of all the RWMDs between a document's keywords and its abstract; this value is then scaled to account for the number of keywords in a document.

Data

This text uses a small corpus of neuroethics articles to demonstrate the utility of the KDR for comparing author-provided keywords to Scopus-assigned keywords (i.e., index terms). The data used come from Scopus, which is an ideal bibliographic database because it has good journal coverage in both the humanities and health-related sciences and indexes more journals than other databases, such as the Web of Science (Falagas, Pitsouni, Malietzis & Pappas, 2008). A keyword-based query was used to obtain publication records from Scopus because this approach has proven successful for neuroethics (Leefmann, Levallois & Hildt, 2016). I searched Scopus for any articles published in English that contain "neuroethic*" in the title, abstract, and/or keyword fields. After excluding articles with missing abstracts and/or keywords, there are 727 publications. In calculating the distance from keywords to documents, this paper uses abstracts—rather than full article text—to represent a document.

Results

Table 1 presents the descriptive statistics on both sets of keywords. It is notable that, although the numerator and denominator are higher in the index terms KDR than in the author keywords KDR, the numerator is proportionally much higher in the index terms KDR. The mean and median KDRs are much lower for author keywords. Note that values below 1 indicate that a document's cumulative keyword-to-abstract distance is greater than the document's keyword-to-keyword cumulative RWMD; values greater than 1 indicate that a document's cumulative keyword-to-abstract distance is less than the document's cumulative keyword-to-keyword RWMD. In other words, KDR < 1 indicates that keywords are more similar to each other than to the abstract, and KDR > 1 indicates that keywords are less similar to each other than to the abstract. Table 1 suggests that authors tend to choose keywords that are more similar to each other while Scopus assigns index terms that maximize their differences.

Table 1. Descriptive statistics about the KDR and its main components for both author keywords and Scopus index terms

Value	Mean	Med.	SD	Min.	Max.
Author Keywords					
Numerator	20.3	13.5	21.4	1.06	184
Denominator	22.6	23.1	23.1	0.00	200
KDR	.881	.896	.0997	.381	1.16
Index Terms					
Numerator	406	243	562	.456	8107
Denominator	71.7	58.0	57.4	0.00	549
KDR	4.82	3.62	5.74	.0822	106

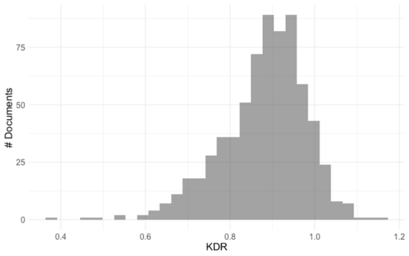
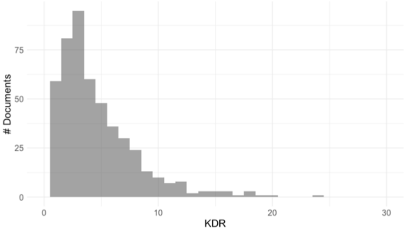


Figure 1. KDR distribution based upon author keywords

Aside from these descriptive statistics, it is instructive to visualize the distribution of these KDRs calculated from different types of keyword. Figure 1 presents the KDR distribution when calculated based upon author keywords, and Figure 2 presents the KDR distribution for index terms. The difference in the *x*-axes was to be expected from Table 1. However, the difference in distribution shape is striking. The author keyword KDR distribution is somewhat normally distributed around 1, while the index term KDR distribution is dramatically skewed to the right.



This figure includes Scopus index terms and excludes 125 observations for presentation purposes.

Figure 2. KDR distribution based upon Scopus index terms

Ongoing analyses (not included) examine the impact of potential limitations by increasing sample size and looking at less interdisciplinary fields/topics as well as other data sources (e.g., Web of Science). Future research should investigate specific mechanisms linking keyword selection practices to KDR distributions, how KDR differences are associated with perception of keyword accuracy and utility, alternative embeddings for calculating distances, and the use of full article text rather than abstracts.

Acknowledgments

This material is based upon work supported by the U.S. National Science Foundation under Grant Number 1939965.

References

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A. & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB Journal*, 22(2), 338-342.

Kusner, M. J., Sun, Y., Kolkin, N. I. & Weinberger, K. Q. (2015). From word embeddings to document distances. *32nd International Conference on Machine Learning*, 37, 957-966.

Leefmann, J., Levallois, C. & Hildt, E. (2016). Neuroethics 1995–2012. A Bibliometric Analysis of the Guiding Themes of an Emerging Research Field. *Frontiers in Human Neuroscience*, 10(336), 1-19.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1-12.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y & Kanakia, A. (2020). Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, 1(1), 396-413.

Ranking Universities via Clustering

George Stoupas[1], Antonis Sidiropoulos[2], Dimitrios Katsaros[3] and Yannis Manolopoulos[1,4]

{grgstoupas,manolopo}@csd.auth.gr, asidirop@ihu.gr, dkatsar@uth.gr, yannis.manolopoulos@ouc.ac.cy

[1]Aristotle University of Thessaloniki, 54124 Thessaloniki (Greece)
[2]International Hellenic University, 57001 Thessaloniki (Greece)
[3]University of Thessaly, 38221 Volos (Greece)
[4]Open University of Cyprus, 2220 Nicosia (Cyprus)

Introduction

University rankings have been approached by academicians, politicians, journalists, and policy makers, even though it is surrounded with scepticism (Manolopoulos & Katsaros, 2017). University rankings are especially conspicuous for the top universities (Angelis, Bassiliades & Manolopoulos, 2019).

All university rankings basically agree on the order of the top-20 or top-30 universities, with major disagreements appearing thereafter. However, one could still wonder whether there is *really* a *serious* performance or prestige discrepancy between universities e.g., either in the range [400-500] or even in the ranges [2-5] or [6-15].

Problem description

Following the reasoning that there is small justification in providing absolute and rigid ranked lists of universities, but it is more meaningful to provide ranked sets, we will use clustering methodologies to achieve our goal. So, the aim of the present study is to use (and evaluate the appropriateness of) popular clustering algorithms to develop university rankings with the defining characteristic that competing universities are "organized" into sets where ranking/ordering is imposed only among the sets, whereas the elements within any set are considered unordered. This concept comprises a departure from existing methodologies where – at least for the first, say, 150 positions – there is a *strict* ranking/ ordering among universities, i.e., total order, whereas we seek for an approach developing partially ordered sets of universities (posets).

Dataset

We choose to work with the National Taiwan University Ranking (NTU) (http://nturanking.csti.tw), founded by the Higher Education Evaluation and Accreditation Council of Taiwan (HEEACT), because it is based exclusively on verifiable research performance indicators. NTU ranking is based on eight features categorized into three categories: *research productivity*, *research impact* and *research excellence* with weights 25%, 35% and 40%, respectively. We work with the top-500 universities of the NTU list.

Experimentation

We used the Weka software (Witten, Frank & Hall, 2011), which offers machine learning libraries, including many algorithms for clustering. For our experiments, we choose the most popular representative of each of three well-known families of clustering algorithms, namely:

- Expectation Maximization (EM) (model-based algorithm)
- DBSCAN (density-based algorithm)
- *k*-means (center-based algorithm)

Our experimental methodology was the following: First, we used the EM-algorithm with the default Weka values, and got 12 clusters. Second, we examined DBSCAN with its default Weka values, except that we set *minpoints*=1, so that we do not lose any outlier, plus we varied parameter *e* with step 0.01 to get the number of clusters. By using the Elbow method (Thorndike, 1953), we came up with 43 clusters. Using the previous findings, we performed a set of experiments on the 3 algorithms for 12 and 43 clusters. From Table 1 we see that DBSCAN does not provide any useful insight for the particular dataset, whereas EM and *k*-means give similar results in terms of the maximum cluster size and the number of singleton clusters.

Table 1. Statistics of the examined algorithms.

	#clusters	max cluster size	#singleton clusters
DBSCAN	12	488	10
	43	430	40
EM	12	90	1
	43	33	2
k-means	12	99	1
	43	23	1

To rank clusters, we assign to each cluster a value equal to the sum of the median values of each of the 8 features mentioned in the "Dataset section". Thus, we order the clusters from 1 (the highest quality) to 12 (or 43).

Table 2 shows the size of each cluster for each of the three algorithms, with EM and *k*-means clusterings bearing similarity in terms of clusters' cardinality.

Table 2. Size of the 12 clusters per algorithm.

	DBSCAN	EM	*k*-means
1	1	1	1
2	1	17	16
3	1	42	28
4	1	34	13
5	1	48	14
6	1	33	55
7	1	44	29
8	2	49	77
9	1	42	41
10	1	90	99
11	1	70	72
12	488	30	55

In Table 2, Harvard is always the (singleton) cluster #1. Cluster #2 of *k*-means and EM consists of: Berkeley, Cambridge, Columbia, Imperial College, Johns Hopkins, Michigan/Ann Arbor, MIT, Oxford, Pennsylvania, Stanford, Toronto, UCollege London, UCLA, UCSD, UCSF and Washington/Seattle. In addition, EM's cluster #2 includes Tsinghua. On the other hand, DBSCAN is much closer to the existing methodologies of university rankings.

Although NTU provides an ordered list, there are tie cases that can be considered as ranked clusters; thus, NTU rank can be viewed as a set of 406 clusters. Table 3 shows the Rand Index (RI) (Rand, 1971) for each pair of clustering algorithms as well as with respect to the NTU ranking for 12 (in parenthesis, for 43) clusters.

Table 3. Rand Index for 12 (43) clusters.

	EM	*k*-means	NTU
DBSCAN	0.147 (0.284)	0.167 (0.280)	0.048 (0.261)
EM	--	0.878 (0.973)	0.890 (0.971)
k-means	--	--	0.878 (0.973)

Oppositely, we view each cluster of the EM or *k*-means algorithm as a list of equal performing universities, ordered according to the position in the NTU list. Thus, we apply the Spearman rank correlation coefficient (ρ) for all pairs of algorithms in Table 3, getting Table 4.

Table 4. Spearman ρ for 12 (43) clusters.

	EM	*k*-means	NTU
DBSCAN	0.230 (0.509)	0.239 (0.500)	0.242 (0.503)
EM	--	0.905 (0.968)	0.960 (0.978)
k-means	--	--	0.896 (0.973)

It seems that *k*-means is the most appropriate clustering algorithm among the three to carry out our goal. Using this algorithm, we investigated the following question which might be crucial for university administrators: Is there any particular feature or set of features (among the eight ones examined) that affect the clustering the most, and which are the methodologies to discover it/them?

Weka has already built-in support for such kind of questions. In Table 5, we see that for the *k*-means, the *average number of citations in the last 11 years* (AveCit, cell in grey) is the most important and affects the ranking dramatically; the rest of the features present very similar rankings with the case when all the 8 features are used. For instance, the features HiCit (*number of citations in the last 11 years*) and CurCit (*number of citations in the last 2 years*) do not affect the ranking as their Spearman (Rand Index) value is close 1.

Table 5. Spearman *ρ* (Rand Index) for feature elimination in the *k*-means.

	8 features without HiCit	8 features without CurCit	8 features without AveCit
Full set of 8 features	0.9671 (0.9834)	0.9568 (0.9396)	0.5190 (0.8519)

Conclusions

We argue for the representation of university rankings in the form of ordered clusters. In the present work, we clustered universities from the NTU ranking with DBSCAN, EM and *k*-means, the last one being the most appropriate. Furthermore, looking at the clusters produced, with the exception of the top (singleton) cluster, the second position is occupied by 16 (17) universities, and based on the concepts developed here, these universities are performing equally well, and thus they can be ranked into the same position.

References

Angelis, L., Bassiliades, N. & Manolopoulos Y. (2019). On the necessity of multiple university rankings. *Collnet Journal of Scientometrics & Information Management*, 13(1), 11-36.

Manolopoulos, Y. & Katsaros, D. (2017). Metrics and rankings: Myths and fallacies. In *Revised selected papers, 18th DAMDID Conference*, pp.1-16, Springer.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *JASIST*, 66(336), pp. 846-850.

Thorndike, R. (1953). Who belongs in the family?, *Psychometrika,* 18(4), 267-276.

Witten, I., Frank, E., Hall, M. & Pal, C. (2011). *Data mining: Practical machine learning tools and techniques*, 3rd edition, Morgan Kaufmann.

RCC: NSFC's new review mechanism: its missions and challenges

Shiyan Tang[1], Weiwei Li[2] , Quan Liu[3] and Yuxian Liu[4]

[1]*shiyan.tang@ucl.ac.uk*
University College London, School of Management, Canary Wharf, London (UK)

[2]*w.li@rug.nl*
University of Groningen, Campus Fryslân, Center for Internationalisation of Education, Wirdumerdijk 34, 8911 CE Leeuwarden (the Netherlands)

[3]*lqmmbb@126.com*
Nanjing Institute of Huawei Technology Co., Ltd, Ruanjian Avenue 101, Nanjing, Jiangsu Province (China)

[4]*yxliu@tongji.edu.cn*
Tongji University, Tongji University Library, Siping Road 1239, Shanghai (China)

NSFC's reform and the new review mechanism

A Reform Plan on the Funding System was approved by the 8th NSFC Committee in 2018. According to the roadmap, in the next 5 to 10 years, three major tasks will be accomplished, namely identifying funding categories, improving evaluation mechanisms, and optimizing the layout of research areas. Improving the evaluation mechanism is a key step to guarantee the success of this reform. A category-specific, accurate, fair, and effective evaluation mechanism is desired to ensure that original ideas of the best scientific merit are timely supported. To achieve this task, the NSFC Committee set up a peer-review mechanism featuring Responsibility, Credit, and Contribution (RCC), based on the active participation of scientists to improve accuracy, fairness, and performance in supporting basic research. In 2019, the NSFC began to experiment with the RCC review mechanism in its review processes.

RCC's mission

Before the RCC, NSFC's review system just had one mission, namely selecting proposals to fund. However, the reform assigns RCC new missions, namely guiding the applicants to improve their proposals and funding the proposals that really belong to identified funding categories.

RCC's challenges

The new missions make the review system face at least two challenges.

The black box of peer review and review biases

Peer review is a fundamental procedure for the NSFC to decide who receive financial support for scientific research. The NSFC adopted a single-blind review system. Reviewers know applicants and their team members, but applicants don't know the names of the reviewers. The NSFC asked reviewers to provide scientific comments, rate the proposals and recommend which proposals deserve to be supported. Based on these rates and recommendations, the NSFC makes a shortlist for a panel discussion and select proposals to be funded. The panel discussion is confidential: whoever lets out any information about the discussion will be punished.

The scientific comments are reported to applicants by the NSFC after the review process is finished. However, they are often biased. Even if there are no prejudices, misunderstandings, and gaps in knowledge, it is hard to arrive at identical evaluations because human judgment can never be completely free from cognitive style, personal views, and social-cultural values. Due to different professional perspectives, review experts often have blind spots and different review focuses. Their views on a project are always disciplinary- and culture-based. So no one should be surprised that peer review is often biased and inefficient.

The selection process is a black box. Which proposals are selected for the panel discussion and how the panel discussion selects the proposals to fund from the biased scientific comments stay unknown. The NSFC doesn't make any efforts to judge whether comments are biased or not. Scientific comments, which are supposed to improve the applicants' research level, are often so confusing that applicants cannot receive any useful guidance.

Categories and the relation between scientific comments

In the reform plan, the NSFC is going to identify four funding categories based on the attributes of scientific problems: creative and timely ideas, research focusing on frontiers of science in unique ways, application-driven basic research, and transdisciplinary leading-edge research. According to the characteristics of these categories, the NSFC made concrete questions and indicators to ask reviewers to answer.

These four funding categories actually locate where the scientific problems are. The NSFC believes that clear identification of funding categories will not only help to improve the quality of applications but also make the evaluation process more efficient and fairer. However, in reality, most reviewers didn't follow the NSFC's instructions. For example, one applicant submitted a proposal in the category of transdisciplinary leading-edge research in 2020. The NSFC forwarded five reviewers' comments to him. Only one reviewer followed the NSFC's instruction and answered all questions that the NSFC wants reviewers to answer. Two of the reviewers asked why the scientific problems are not from one single discipline. One of the reviewers thought that very common concepts in other disciplines are just somethings fabricated in the mind of the applicant.

Under such circumstances, it will be very important for the review system to adopt a suitable method to select the proposals that have the proper characteristics. Normally the review system implements a majority and hierarchy rule. After scientific problems are classified by their characteristics, it is of the utmost importance that reviewers follow the attributions of scientific problems in their comments. Since review biases result from different professional perspectives, blind spots and different review focus, we may expect that combination of different review focuses from different professional perspectives according to their logic may reduce these problems. The relations between the comments of different reviewers must be taken into consideration to identify the scientific problems that have the specific attributions.

The solutions to these challenges: Responsibility+Credit+Contribution enhanced with open and "trust but verify with evidence"

Opening the black box of peer review is a solution to solve the challenges. Many scholars inclined to support openness in peer review, believe that open and transparent reviews lead to more constructive reviews and criticism would be more reasonable (Cosgrove & Flintoft, 2017). More importantly, responsible reviewers who contributed constructive comments should be able to claim credit, which will automatically improve the review process (Moussian, 2016).

Also reducing and correcting review biases is a way to solve the challenges. The NSFC should design its review procedure to make it possible for the reviewers on same proposals to reach an agreement so that the applicants can follow the scientific comments to improve their proposals and research levels. In the current review procedure, reviewers are prohibited to communicate with each other in the communication-based review stage, which also prevent the reviewers to reach an agreement.

Making peer review scientific is fundamental to solve all these challenges. Rennie (2016) described how the field of clinical science developed evidence-based lists of items to be included in review reporting and make sure journals, reviewers, authors, manuscript editors and copy editors uphold the standards. The NSFC should develop evidence-based lists of review items and ensure that all stake-holders uphold the standards. The NSFC only rely on the rate and recommendations from reviewers to select the proposals to fund. The rate and recommendation should be trusted, but must be verified by their evidence in the selection process.

Moreover, peer review must be thoroughly studied. Actually, outside China, a data-sharing infrastructure has been proposed to enable systematic research on peer review. To strengthen institutional cooperation with data sharing on different review formats, a group of science and technology scholars, publishing professionals and funders formed a collaboration called PEERE, funded by COST (the European Cooperation in Science and Technology). In 2017, a PEERE protocol for sharing data on the peer-review process was released, which carefully considers ethics, responsible management, data protection, and privacy (Squazzoni et al., 2020). This infrastructure intend to help open the black box of peer review without offending the reviewers, at the same time, make it possible to study and verify the peer review.

All in all, in the era of funding-oriented reforms based on the attributes of scientific issues, even though the 'Responsibility + Credit + Contribution (RCC)' peer-review system are making the great efforts to identify scientific problems as well as to guide the applicants to improve their proposals and research levels, the challenges need to be taken seriously. Solving this challenges requires the wisdom of all scientific and technological circles. We should not only create an infrastructure to facilitate the study of peer review, but also supervise the review process with "trust but verify" attitude to ensure the peer review scientific.

Acknowledgments

The authors thank Ronald Rousseau, Liexun Yang, Yishan Wu, Cong Cao, Tiecheng Guo and Liying Yang for helpful discussions.

References

Cosgrove, A. & Flintoft, L. (2017). Trialing transparent peer review. *Genome Biology*, 1*8*(1), 17-18.

Moussian, B. (2016). Taking peer review seriously. *EMBO Reports*, 17(5), 617-617.

Rennie, D. (2016). Make peer review scientific, *Nature*, 535(7610), 31-33.

Squazzoni, F., Ahrweiler, P., Barros, T., Bianchi, F., Birukou, A., Blom, H. J. J., … Willis, M. (2020). Unlock ways to share data on peer review. *Nature*, 578(7796), 512-514.

Broad shouldered - publication formats served by doctoral students

Jakob Tesch[1]

[1] *tesch@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW), Department Research System and Science Dynamics, Schützenstraße 6a, 10117 Berlin (Germany)

Purpose of the study

Studies of publication output from doctoral training on a national level covering all disciplines remain rarities. Larivière (2012) covers a Canadian province and Nettles and Millett (2006) as well as Enders and Bornmann (2001) selected disciplines only. This exploratory study takes a closer look at the variety of publication formats served during the PhD based on dissertations thus exploring the usefulness of this data source for studying formats and extent of doctoral publishing.

Data and Methods

This study uses a 5% random sample of German PhDs based on cleaned dissertation meta-data collected for the 2014 cohort (n = 1,381), which was downloaded from the German National Library (DNB) in 2019 (Tesch et al., 2021).

This study differentiates theses published by non-commercial- (NCPs) and commercial publishers (CPs, 18% of the sample). 54% of the sample are Electronic Theses and Dissertations (ETDs) the subset of the NCPs that are available in open access format (OA) and together with a few OA-CPs (3%) are thus fully accessible. The 41% of theses where Tables of Contents (TOC) but not the full text was available are termed "partially accessible theses." Only 2% of the sample were not accessible because the TOC was not available.

Identification of published work

Published work is a binary variable for presentation of results and is coded "yes" if the thesis contains references to at least one of the following publication formats authored by the PhDs: published-, in press- or accepted journal articles, published proceeding papers and contributions to edited volumes. Manual coding identified published work based on a coding scheme. Partially accessible thesis count as containing published work if the thesis contains a section entitled "own Publications" or similar or if publications related to the thesis title could be found using Google Scholar. 58 Thesis from Law were excluded because common search engines do not index publication formats typical for this discipline. Results report single disciplines if the Chi² test results shows the share of publishing PhDs to differ significantly from their official broader disciplinary grouping (e.g. Psychology in the case of Social Sciences).

Findings

Table 1 shows the different formats of the theses as well as the share containing published work in total and with at least one journal article authored by the PhD.

Table 1. Formats of German dissertations and published work (n=1,323)

	Publication format of the thesis		*Published Work*	*Journal Article*
Discipline	*Share ETDs in %*	*Share CPs in %*	*Share "yes" in %*	*Share with at least 1 in %*
Arts & Humanities	23%	69%	22%	13%
Social Sciences	37%	41%	39%	31%
Psychology	48%	8%	60%	40%
Natural Sciences	69%	5%	72%	66%
Mathematics	82%	0%	36%	27%
Medical Sciences	53%	1%	34%	29%
Agricultural- & Food Science	66%	3%	84%	81%
Engineering	51%	26%	59%	37%
Computer Science	94%	6%	94%	58%

Theses based on a series of papers are difficult to distinguish but they are unlikely to appear in CPs due to copyright issues. Yet in 16% of the CPs overall and 60% of the fully accessible CPs references to published work could be identified- most of them contributions to edited volumes.

The differences between the share of published work and PhDs with at least one journal article in Table 1 hint to the importance of publication formats other than journal articles in the discipline. For example, while coding identified published work for 94% of PhDs in Computer Science only 58% of these PhDs published at least one journal article.

The share of PhDs with at least one article is higher here than reported in Larivière (2012) for most disciplines which is plausible because here no restriction to WoS was made but it is lower in the

Medical Sciences. This may result from the peculiarities of the German system which does not know of a MD degree and the high share of students from Medicine graduating with a Dr. med (Konsortium Bundesbericht wissenschaftlicher Nachwuchs, 2017).

Number of publications in different publication formats

Theses containing published work differ in terms of number and type of this work. Figure 1 underlines the importance of proceedings in Engineering Sciences and even more so in Computer Sciences. The mean of journal articles in the Natural Sciences is close to three, a number often required for a thesis as a series of papers. Natural- and Medical Sciences are more homogenous in terms of relevant publication formats, which are articles and proceedings while all other disciplines address a broader range of formats.

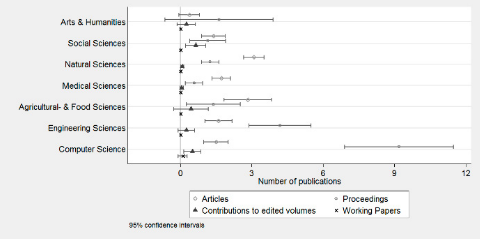


Figure 1. Number of publications in fully accessible PhD theses containing published work by discipline and publication format (n=528)

Discussion

The purpose of this study was to shed more light on the different publication formats served by PhDs during their studies. Based on dissertation metadata and full texts of dissertations this study finds results comparable to earlier studies in terms of the share of published work and PhDs with at least one journal article published or accepted prior to submission of the theses. Yet at the same time, this study highlights the high share of theses published in CPs in Social Sciences and Humanities and shows that these theses also contain published work by the PhDs. Regarding the number of publications in different formats Medical- and Natural Sciences are more homogenous as they focus on publishing in journals and proceedings. All other disciplines serve a broader range of different publication formats including edited volumes and working papers to a lesser extent.

The approach based on dissertations presented herein reports larger shares of PhDs with journal articles than for example Larivière (2012) while at the same time providing a more complete picture of the publication formats served. Nevertheless, this study may not fully cover publications if they are unrelated to the thesis or appear way before the PhD. Thus, future research should compare results for doctoral publishing based on matching to WoS with that on dissertations to assess the amount undercoverage for both approaches.

Conclusions

A large share of PhDs contributes significantly to research output in a disciplinary specific form. Dissertation metadata and full texts are a promising data source for studying research output (co-)authored by PhDs. However, given the low shares of ETDs in Humanities, Social Sciences and the high amount of manual work associated with coding, a combination of dissertation metadata and matching to WoS seems a promising avenue for investigating doctoral publishing in the future.

Acknowledgments

This research is part of the RISIS2 project funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091. The author likes to thank Simrit Rattan for her valuable help with the manual coding.

References

Enders, J. & Bornmann, L. (2001). *Karriere mit Doktortitel?* Frankfurt a.M./New York: Campus.

Konsortium Bundesbericht wissenschaftlicher Nachwuchs. (2017). *Bundesbericht wissenschaftlicher Nachwuchs 2017*. Bertelsmann W. Verlag. Retrieved from https://www.buwin.de/dateien/buwin-2017.pdf

Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. *Scientometrics*, 90(2), 463-481.

Nettles, T. M. & Millett, C. M. (2006). *Three Magic Letters: Getting to Ph.D.* Baltimore, Maryland: John Hopkins University Press.

Tesch, J., Iversen, E. J., Skålholt, A., Franssen, T., Honk, J. van, Cañibano-Sanchez, C., Perruchas, F., et al. (2021). *Manual & Documentation of Doctoral Degree and Career Dataset (DDC) – first version*. Zenodo. Retrieved February 13, 2021, from https://zenodo.org/record/4446431

Words Connecting Scientific Communities
The case of ‘argumentation’

Natalija Todorovic[1] and Benedetto Lepori[2]

{[1] todorn, [2] blepori}@usi.ch
Università della Svizzera Italiana, Via Buffi 13, 6900 Lugano (Switzerland)

Introduction

Science is a communicative social creation, whose main activities consist of sharing and transmission of scientific contents, with a goal of clarification and comprehension of concepts, values, and argumentations (Merton, 1996). The cognitive concepts that are being shared across communities might carry ambiguity, given that their interpretation might differ according to the context. Mc Mahan and Evans (2018) suggest that ambiguity acts as boundary object between different scientific communities, which drives communication and engagement, a consequence of what might be new interdisciplinary community. Ambiguity can be an important strategy used by scientists that helps accepting the same message by different communities, even if the interpretation is different (Ceccarelli, 2001). The ambiguous words thus become a connector of two otherwise separate communities, possibly enabling interdisciplinarity.

Rafols and Meyer suggested that interdisciplinarity is characterized by two dimensions: disciplinary diversity and network coherence. Diversity, that comes from differences in cognitive rooting, is measured as “heterogeneity of a bibliometric set from predefined categories”, while coherence represents increasing interconnectedness in the social structure.

In this study, we look at the communities using the keyword ‘argumentation’ to see whether and how the two mechanisms suggested by Rafols and Meyer (2010) together enable interdisciplinary research. We are interested in understanding the interdisciplinarity enabled by ambiguous language.

Argumentation studies are an area of inquiry often seen as having an important interdisciplinary appeal (Reed & Koszowy, 2011; Van Eemeren et al., 2014). There is an increasing usage of term “argumentation” in scientific literature with subdisciplines having different disciplinary rooting (ranging from philosophy to informatics), different approaches, and, possibly, different understanding of core concepts, argumentation being one of them. Nevertheless, we think that the term that is common across different disciplines drives communication and facilitate coordination among them.

Despite claims in literature that there is increasing integration of argumentation studies filed, we know little about its structure and we specifically miss: a) the extent of integration between different areas or subcommunities, b) the extent of diversity in the community and how concepts and terms are shared and interpreted across sub-communities. To tackle these gaps, we rely on the concept of interdisciplinarity and on the insights from science studies. Even so authors suggest the argumentation becomes interdisciplinary field (Keith & Rehg, 2008; Reed & Koszowy, 2011; van Eemeren et al., 2013; Walton, Reed, & Macagno, 2008), their work seems to be based on perceptions and we did not encounter a systematic analysis based on empirics. What also seems to be missing in literature is rooting of interdisciplinarity through integration and diversity. With this study we are aiming at closing these gaps.

Data Collection

The data was retrieved from Scopus database using two recursive queries. We first queried for the term ‘argumentation’ in abstract, title and/or keywords of the publications (in August 2020), focusing on results published from 2000 until 2019, published in English. After consulting experts from the field we performed additional querying to include results with ‘argument scheme’, ‘argument mining’, ‘argument interchange’, ‘argumentative discourse’, ‘argumentative dialogue’ and ‘argumentative strategies’ in title, abstract or keywords. After cleaning the data our final data set consists of 11178 publications.

Methods

We applied scientometric techniques, text mining and map visualization offered by VOSviewer software (van Eck & Waltman, 2010). The goal is to identify the main topics around which the communities formed using co-occurrence map analysis. Through analysis of scientific map, we reveal the structure of scientific community, that combined with expert knowledge from the community completes comprehensive picture of broad interdisciplinary community in question (Chandra, 2018). To execute topic mapping, we used VOSviewer’s natural language processing functionality. First, we crated thesaurus to cancel general ,“noise” and ambiguous terms (ex. Tree, note, goal) and to replace all plural forms of terms with their singular form. We chose to include in our analysis only terms that appeared more than 7 times.

The relatedness of items on map is determined on the number of documents in which they appear together (van Eck & Waltman, 2007).

Results

The world 'argumentation' is used in almost all subject categories, although it has the most weight in social sciences (27%), computer science (23%), and arts and humanities (18%). This wide usage of the term is reflected in journals and other sources of publications, as well.

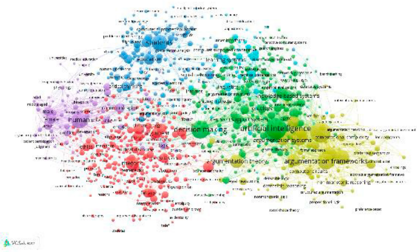


Figure 1. Keywords co-occurrence map

We examined the keywords in each cluster and named the clusters as following: *a) discourse analysis, b) interpersonal communication, c) argumentation in science and education, d) computer science, e) argumentation frameworks, f) natural language processing*. The first three clusters belong to social sciences and humanities, while the latter three are computer science. Some terms (like 'decision making', 'argumentation theory', etc.) are acting as point of connection representing the idea that language is not a simple collection of words, and that logical structure is important for transmitting a message and achieving the conclusion as a core of argumentation studies (Willard, 2003).

Conclusion

Starting from one word shared across communities we can analyse not only a single intellectual tradition it comes from, but also its declensions in other contexts. To do so we must understand both cognitive, social, and institutional structures of community and how these mutually connect. Starting from term 'argumentation' we discovered that it is used in some strictly disciplinary areas like science studies, that use this word as a complement for their subject of study. In some areas, such as rhetoric, 'argumentation' is primary object of inquiry. Other areas, like artificial intelligence, apply models and logic from argumentation studies in IT contexts. Thus, argumentation studies have relatively small and partially interdisciplinary core, with large, applied periphery. These findings might take us to further reflect on how scientific fields are structured.

References

Ceccarelli, L. (2001). *Shaping science with rhetoric: The cases of dobzhansky, schrodinger, and wilson* University of Chicago Press.

Chandra, Y. (2018). Mapping the evolution of entrepreneurship as a field of research (1990–2013): A scientometric analysis. *Plos One,* 13(1) Retrieved from https://doi.org/10.1371/journal.pone.0190228

Keith, W. & Rehg, W. (2008). Argumentation in science: The cross-fertilization of argumentation theory and science studies. *The Handbook of Science and Technology Studies*, 211-239.

McMahan, P. & Evans, J. (2018). Ambiguity and engagement. *American Journal of Sociology,* 124(3), 860-912.

Merton, R. K. (1996). In Sztompka P. (Ed.), *On social structure and science*. Chicago/London: The University of Chicago Press.

Rafols, I. & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, 82(2), 263-287.

Reed, C. & Koszowy, M. (2011). The development of argument and computation and its roots in the Lvov-Warsaw school. *Studies in Logic, Grammar and Rhetoric, Special Issue of the Argumentation Series on Argument and Computation, Ed.Koszowy, M*, 23(36), 15-37.

van Eck, N. J. & Waltman, L. (2007). VOS: A new method for visualizing similarities between objects. Paper presented at the *Advances in Data,* pp. 299-306.

van Eck, N. J. & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538. doi:10.1007/s11192-009-0146-3

van Eemeren, F. H., Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, A. F., Verheij, B. & Wagemans, J. H. M. (2013). Argumentation and artificial intelligence. In F. H. van Eemeren, B. Garssen, B. Verheij, E. C. W. Krabbe, A. F. Snoeck Henkemans & J. H. M. Wagemans (Eds.), *Handbook of argumentation theory* (pp. 1-53). Dordrecht: Springer Netherlands.

Van Eemeren, F. H., Garssen, B., Krabbe, E. C., Henkemans, A. F. S., Verheij, B. & Wagemans, J. H. (2014). *Handbook of argumentation theory*.

Walton, D., Reed, C. & Macagno, F. (2008). *Argumentation schemes* Cambridge University Press.

Willard, C. A. (2003). *A theory of argumentation* University of Alabama Press.

Linking the Web of Science and Scopus to a Publication Repository

Dirk Tunger[1], Roland Erwin Suri[2], Andreas la Roi[3] and David Johann[4]

[1]d.tunger@fz-juelich.de
TH Köln, Institute of Information Management, Claudiusstraße 1, 50678 Köln (Germany)
Forschungszentrum Jülich GmbH, Project Management Jülich, 52425 Jülich (Germany)
External consultant at ETH Library, ETH Zurich, Rämistrasse 101, CH-8092 Zürich (Switzerland)

{[2] roland.suri, [3] andreas.laroi, [4] david.johann}@library.ethz.ch
[3] https://orcid.org/0000-0003-4209-1361
ETH Library, ETH Zurich, Rämistrasse 101, CH-8092 Zürich (Switzerland)

Introduction

It is common practice for universities to collect research publications in publicly accessible databases. As they need a repository of all publications for reporting, the university collections also provide detailed information about the organizational units (e.g., departments). One example of such a repository is the so-called Research Collection (RC) hosted by the library of ETH Zurich. We link the RC with citation statistics from "Web of Science" (WoS) and "Scopus" to be able to prepare detailed bibliometric reports, e.g., for the ETH's controlling department. It is outlined below how this can be achieved and why this is a sensible approach for higher education institution.

Function and strategy of the RC

The RC serves as a single point of access for ETH Zurich staff to document, publish, and archive the scientific results of their research. Thus, it fulfils multiple functions as a publication directory, an open access repository, and a research data repository (Hirschmann, 2018): The RC contains publications authored by ETH staff and serves as a publication platform for *all* types of research results. This implies that staff can also publish research data using the RC. It further serves as a source to produce lists of publication for the academic Annual Reports and the website of the ETH Zurich. It is also an instrument to implement the ETH Zurich's Open Access Policy, because it gives university staff the opportunity to publish scientific papers beyond traditional publication platforms.

Linking the RC to Publication Databases

Many organizational units at the ETH Zurich are interested in providing evidence of and documenting publication achievements, but also in performing bibliometric analyses of their publication success. For example, one common request, e.g., by the ETH controlling department, is to compile a list of all publications in the last five years for a specific university department and to supplement it with current citation numbers. Occasionally, such a request is only the starting point for more in-depth queries for bibliometric evaluations, e.g., in the form of a field normalization.

Commercial publication and citation databases, such as the WoS and Scopus, do not provide enough detail to validly assign publications to smaller organizational units within an institution. Thus, relying on these databases alone to process such a request is insufficient. This is where resources such as the RC provide valuable additional information. While the RC itself does not contain any citation information about the publications covered by the system, it remains the only central source that maps scientific outputs of all organizational units of the ETH Zurich. As such, it can serve as a link between publication outputs and citation statistics of the WoS and Scopus. To process requests like those described above, the RC are matched with the information of the WoS and Scopus.

Matching the RC with WoS and Scopus

We rely on licenced data provided to the ETH Zurich by the WoS and Scopus. For the WoS, the Core Collection Citation Indexes are used. We consider articles, conference proceeding, reviews, and letters for our analysis presented below.

To establish a connection between publications in the RC and the WoS and Scopus databases, the publication IDs (Accession Numbers (ANs)) are required. Next, the ANs need to be assigned to the publications by the database providers. In the WoS the ANs are commonly called the UT code (field tag UT in the underlying export-spreadsheet); in Scopus EID. In order to extract detailed information about the citation data for the ETH departments, each item's metadata in the RC needs to be supplemented with the ANs of the WoS and Scopus.

Researchers and administrative staff at the ETH submit the metadata of their publications (sometimes also a post-print) to the RC (Hirschmann, 2018). The RC also imports publication metadata using the APIs of the WoS and Scopus. The imports use the publishers' DOIs and ANs for matching the items and to avoid duplicate entries. Metadata imports may contain missing publication items or missing metadata fields.

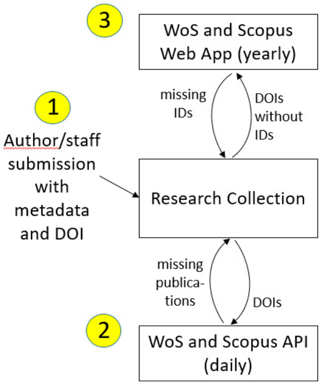

3
WoS and Scopus Web App (yearly)
missing IDs
DOIs without IDs
1
Author/staff submission with metadata and DOI
Research Collection
missing publica-tions
DOIs
2
WoS and Scopus API (daily)

Figure 1. Simplified illustration of data imports to the RC. The data in the RC are provided by authors/ETH staff, WoS, and Scopus in 3 steps.

(1) Metadata that were manually entered by the authors or the ETH are supplemented with the ANs from the WoS and Scopus. This is done daily by matching items using their DOIs. (2) In the same process also the metadata of missing publications are added to the RC. If the DOI of an ETH publication cannot be found in the RC, the metadata are imported from the WoS and Scopus APIs.[1] (3) Processes (1) and (2) provide ANs for only 87.8% (WoS) and 63.1% (Scopus) of the publications (articles, reviews, letters, and conference proceedings) in the RC. Thus, missing WoS and Scopus IDs will be imported once a year. This solution ensures completion of the ANs for over 90% of the publications in the RC (WoS: 95.1%, Scopus: 91.6%). The proportion of papers that do not match include those papers for which no AN/DOI has been deposited and, thus, could not be found in the WoS or Scopus. The reason for the lack of DOI or AN in the RC needs to be further explored. Table 1 summarises the success of the matching procedures.

Table 1. ETH Zurich publications (2013-2019) identified in the RC by AN and DOI.

	***N* of ETH publica-tions**	***N* of ETH publica-tions in RC with AN**	***N* of ETH publica-tions in RC with DOI**	***N* of ETH publica-tions in RC with AN or DOI**
WoS	45,339	39,798 (87.8%)	40,153 (88.6%)	43,105 (95.1%)
Scopus	51,855	32,714 (63.1%)	46,755 (90.2%)	47,524 (91.6%)
	***N* of ETH journal articles**	***N* of ETH journal articles with WoS AN**		
RC	38,157	37,505 (98.2%)		

Final remarks

The RC paints a detailed picture of all organizational units' scientific outputs (publications) at the ETH Zurich. The benefit of the RC is that the WoS IDs of most publications are known for the organizational units at the ETH Zurich, so it is possible to determine citations for each of these units by using the WoS Web App. In addition, the completeness of the metadata is continuously checked and missing publications are added by employing a daily automated matching procedure on the basis of the WoS ANs and DOIs. DOIs can also be used for matching RC metadata with publication databases in a way that ANs are available for over 90% of ETH publications in the RC.

Our approach uses the WoS and Scopus for metadata enrichment. Using other commercial citation databases, such as Dimensions, would likely produce similar results. Future work will reveal how successful these databases are in enhancing the RC metadata. Any HEI institution that wishes to perform citation analyses for individual departments or other institutional subunits can apply the approach presented here, given that they have access to the WoS or Scopus and that an internal repository is available that contains nearly-complete information, such as DOIs or ANs.

References

Hirschmann, B. (2018). Die Research Collection der ETH Zürich. *ABI Technik*, 38(3), 223-233.

[1] Because of limitations in the WoS and Scopus search for metadata retrieval, only items that are correctly attributed to the ETH Zurich by our search queries in WoS and Scopus are included.

Measuring "Science Teaching" in Higher Education - An Empirical Study Of Teaching Content

Carolin Vollenberg[1], Julian Koch[2], Christopher Still[3] and André Coners[4]

{[1] vollenberg.carolin, [2] koch.julian, [4] coners.andre}@fh-swf.de

[3]christopher.still@stud.fh-swf.de
University of Applied Science Südwestfalen, 58095 Hagen, (Germany)

Introduction

Scientific work is the basis of an academic education and implies the systematic analysis of a specific problem to gain assured knowledge (Ellis, 2004). Within scientometrics, however, the form, application, and expression of the mediating mode of "scientific practice" in teaching has not been adequately explored. This paper aims to present a method and analysis of teaching content on the area of "scientific work" in higher education. In this paper, we analyze module handbook data with a novel approach to show what benefits can be generated with this unique database. The resulting research questions are as follows:

1.) Are there discernible differences between universities and universities of applied sciences in relation to the teaching of "scientific work "?

2.) Which of the study programs has the largest proportion of "scientific work" in teaching?

3.) Is "scientific work" taught as an independent teaching unit or as part of the content in other teaching modules?

Methodology

As already mentioned, this work analyses and measures the textual content from module handbooks of universities and universities of applied sciences (hereafter referred to as "U" and "UAS"). This makes it possible to make statements about the extent to which and the form in which scientific work is taught in German university teaching. To be able to carry out the intended research in a meaningful way, a list of search operators first had to be compiled (Han, Kamber, & Pei, 2011). To create such a list, interviews were conducted with five experts in the field of teaching "scientific work" (Creswell & Creswell, 2017). This resulted in a list of suitable word combinations for identifying "scientific work" within the module handbook data. The first step in our entire research process (cf. Figure 1) is to compile the data sources of the U's and UAS's in Germany. For this purpose, we select the three mass study programmes Business Administration (B. A.), Computer Information Systems (C. I.) and Business Information Systems (B. I.). In the following step, the identified websites of the degree programmes are searched for relevant module handbooks with the help of a crawler, downloaded, transformed, extracted, and stored in a database. After the entire crawling process, the transformation process begins by converting the database into a structured text format. This is a necessary step for the subsequent text pre-processing (Feldman & Sanger, 2007). Using a text mining application, we then perform noise reduction in the form of tokenisation, lower case, stop word removal, stemming and lemmatisation (Allahyari et al., 2017). The database then contains structured data, e.g., the name of the U or UAS, the degree program, the module name and description, the specified SWS (semester hours per week) and ECTS (European Credit Transfer System) points per module. This structuring of all module handbooks subsequently enables us to carry out a variety of measurements and analyses, e.g., for comparison between U's and UAS's, between institution types (U's and UAS's), degree programs or module offerings.

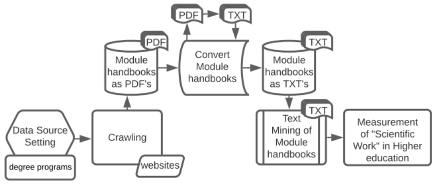

PDF
TXT
PDF
Module handbooks as PDF's
Convert Module handbooks
TXT
Module handbooks as TXT's
Data Source Setting
degree programs
Crawling
websites
TXT
Text Mining of Module handbooks
Measurement of "Scientific Work" in Higher education

Figure 1. Research process

With the help of text-mining algorithms, the module descriptions are searched for the given search operators of the teaching area "scientific work". To answer our research questions, a comparative analysis follows, both at the level of the higher education institutions (hereafter referred to as "HEI") - U's and UAS's - and at the level of the degree programs. We compare whether "scientific work" is anchored as an independent teaching unit or module, or whether it is found as part of the content of an external module, which then also includes other content. Furthermore, the two types of HEI's are compared at degree program level according to the average number of "scientific work"-related content per teaching module. In addition, the proportion of modules with "scientific work"-related content is compared in the chronological semester position of the degree programs.

Results and Discussion

Within the scope of our study, we first measured the number of degree programs with an independent teaching module "scientific work". The proportion of degree programs with an independent module at UAS in Germany is 33.66% on average. This is significantly higher than at the U with a share of 5.94%. Most degree programs (cf. 60.40%) at both types of HEI's do not offer an independent module on "scientific work" at all. The results at the level of the HEI's (U's and UAS's) can also be confirmed at the level of the study programs.

The proportion of individual degree programs at UAS's with a stand-alone module is significantly higher compared to U's. C. I. (7.27%) at U's has the highest proportion of stand-alone modules compared to B. A. (3.96%) and B. I. (6.52%) programs. In comparison, C. I. (20.91 %) has the lowest proportion compared to B. A. (40.59 %) and B. I. (41.30 %) at UAS's. Secondly, we measured the average number of "scientific work" related contents in the teaching modules that are not named as a separate module and thus also cover other content. Figure 2 shows the average number of teaching modules that at least also contain "scientific work" related content, separated by the individual degree programs for the two types of HEI's. This shows that the average number of "scientific work"-related content in the modules is higher at the U's than at the UAS's.

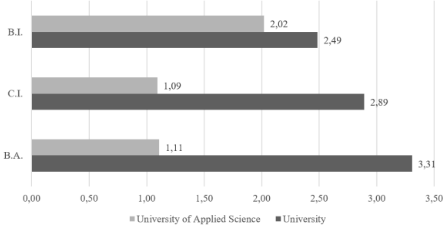


Figure 2. Average number of teaching modules with "scientific work"-related content

In addition, we show the distribution of the proportion of modules with "scientific work"-related content over the period of the (mostly seven-semester) degree programs, differentiated by U's and UAS's (cf. Figure 3). Here it becomes clear that the UAS's place the emphasis on teaching "scientific work"-related content at the beginning of the degree program, while the U's integrate the teaching of "scientific work" into the degree programs without significantly recognizable periods of preference.

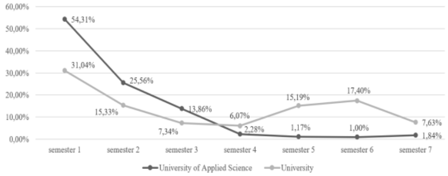


Figure 3. Semester location of modules with "scientific work"-related content

In summary, our analyses show that at UAS's the emphasis is on teaching "scientific work" in stand-alone modules and content of scientific work, especially in the first semesters of the degree programs. In contrast, at U's "scientific work" is not taught independently, but only as a part integrated within thematically different teaching modules.

This contribution to the question of how and to what extent "scientific work" is taught in academic higher education shows that our unique dataset offers enormous opportunities to analyze the content in academic higher education from the perspective of teaching. The paper shows that this approach also has the potential to generate valuable quantifiable results on teaching development, comprehensible thematic characteristics of teaching content, and measurable study trajectories in the future.

References

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Creswell, J. W. & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches:* Sage publications.

Ellis, R. A. (2004). University student approaches to learning science through writing. *International Journal of Science Education,* 26(15), 1835-1853.

Feldman, R. & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data:* Cambridge university press.

Han, J., Kamber, M. & Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, 83-124.

Do UTAUT variables influence the adoption and use of open access scholarly publishing in Kenya? A case study of selected universities

Mercy W. Waithaka[1] and Omwoyo Bosire Onyancha[2]

[1] *mwangechi2013@gmail.com*
University of Nairobi, PO Box 30197-00100 Nairobi (Kenya)
Dept. of Information Science, University of South Africa, PO Box 392, Unisa 0003 (South Africa)

[2] *onyanob@unisa.ac.za*
Dept. of Information Science, University of South Africa, PO Box 392, Unisa 0003 (South Africa)

Introduction

The discourse around open access (OA) in scholarly publishing is an ever-present attraction among institutions, governments and government agencies, publishers and journal vendors, journal indexing services and funders who are increasingly demanding OA scholarship, especially regarding government-funded research. Researchers are particularly interested in the discourse mainly because they are central to knowledge production and sharing through research publications, among other research outputs. Funded researchers, who are the eventual authors of the research publications, are particularly under pressure to publish their research findings in OA platforms. Consequently, many institutions and/or governments have put in place mechanisms to mandate OA for federally funded research (see Kimbrough & Gasaway, 2016; Larivière & Sugimoto, 2018; Bryan & Ozcan, 2020). According to Larivière & Sugimoto (2018) 50 funders and 700 research institutions worldwide, had mandated some form of OA for their funded research by September 2018. Although Kenya does not have an OA policy or registered mandate in the Registry of Open Access Repository Mandates and Policies (ROARMAP) (Chilimo et al., 2018; Waithaka & Onyancha, 2021), the country is listed as one of the countries that have witnessed an increase in the number of OA publications in the recent past. However, the reasons and factors that could be influencing the publication trends and patterns are unknown. Using the Unified Theory of Acceptance and Use of Technology (UTAUT) as a theoretical lens, this study sought to examine the extent to which the UTAUT variables influence the adoption and use of open access scholarly publishing (OASP) in selected universities in Kenya.

Research Methodology

The study adopted a quantitative research approach, with the survey being the research design. The target population comprised 3009 teaching staff from three universities that were selected from the Nairobi County and its vicinity; that is the University of Nairobi [UoN] (n = 1472), Jomo Kenyatta University of Agriculture and Technology [JKUAT] (n = 643) and Kenyatta University [KU] (n = 894). The sample size of 341 was drawn using the sample size determination formula and table proposed by Krejecie and Morgan (1970), at 95% confidence level and .05% confidence interval (margin error). A self-administered questionnaire consisting of several items representing UTAUT variables was used to collect data. Out of the 284 questionnaires distributed, 341 were returned but only 281 were found to be usable in the study. The study used path regression analysis using intention to use and the actual use of OASP as outcome variables regressed against six influencers: (i) attitude; (ii) performance expectancy; (iii) effort expectancy; (iv) social influence; (v) Internet skills or efficacy; and (vi) facilitating conditions, and four demographic factors used as moderating factors: (i) age; (ii) experience; (iii) gender; and (iv) rank. The path regression analysis uses multiple regression equations that specify explanatory variables and the dependent variable at every stage.

Results and discussion

The model estimation yielded a value of adjusted R^2 as 0.994, an indication that there was variation of 99.4% on the dependent variable. The model worked well in attitude or perception, performance expectancy, effort expectancy, social (peer) influence, facilitating conditions and Internet skills or efficacy and intentions to use OA scholarly publishing. In addition, the adjusted multiple coefficient of determination of 0.993 implied a high joint impact of the explanatory variables. Figure 1 provides a summary of the findings based on the regression modelling of the factors influencing OASP in the universities investigated in the study. the model shows that facilitating conditions and efficacy (Internet skills) significantly influence the adoption of OA scholarly publishing while attitude, performance expectancy and social influence significantly influence intention to use OASP in the selected public universities in Kenya. Facilitating conditions emerge as main predictor variable that greatly influence the actual use of OASP among the researchers with the highest beta value of B = 0.521, p < 0.05, with the Internet skills/efficacy recording a value of B = 0.323, p < 0.05. Attitude (B = -0.045, p > 0.05) and social influence (B = 0.120, p > 0.05) have insignificant influence on researchers' OASP adoption while performance expectancy and effort expectance had no influence at all on the same. On the other hand, attitude (B = 0.238, p < 0.05), performance expectancy (B = 0.227, p < 0.05) and social influence (B = 0.125, p < 0.05) significantly influenced researchers' intention to use OASP while effort expectancy and efficacy insignificantly influenced the same. In terms of the effect of demographic variables as moderators of independent variables, age, experience and gender had significant moderating effect on the independent variables. Each moderating variable had varied effects on the independent variables in the latter's influence on intent to use and actual use of OASP. Gender had the most moderating effect, if the results of its significant coefficients are anything to go by. The rank or position of the researchers did not produce any significant beta coefficients, thereby signalling its insignificant moderating effect on all the independent variables. A number of scholars have observed similar patterns in different countries (e.g. Dulle, Minishi-Majanja & Cloete, 2010; Singeh, Abrizah, & Karim, 2013; Bashorun et al., 2016).

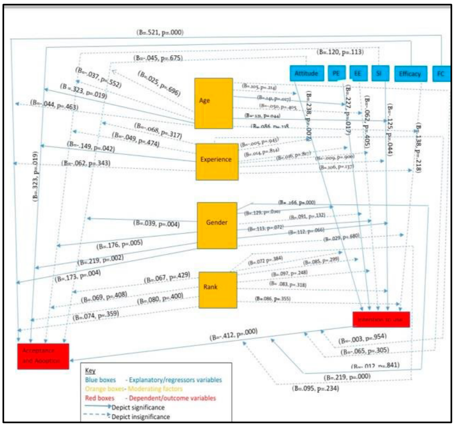


Figure 1: Theoretical model for adoption and use of OASP

Conclusions

The study concludes that the UTAUT variables significantly influence adoption and use of OASP by researchers in universities in Kenya. However, the extent to which each variable influences the adoption and use of OASP varies. The facilitating conditions (e.g., management/financial support; strong and dependable information technology infrastructure; content availability; strong mandates; training capacity; computer and networking skills; advocacy; Internet proficiency and the existence of OA channels and outlets) play a greater role than the other variables in influencing the researchers' intent and actual use of OASP. The Internet proficiency and self-efficacy featured as the second most important influencers or predictors of the adoption and use of OASP. While some independent variables such as facilitating conditions and efficacy have a direct and significant influence on the adoption and use of OASP, the attitude, performance expectancy and social influence significantly influence the intent to use OASP, which in turn significantly influences the adoption of OASP.

Acknowledgement

This paper is an extract from a PhD thesis compiled by the first author under the supervision of the second author.

References

Bashorun, M. T., Jain, P. & Sebina, P. M., & Kalusopa, T. (2016). Determinants of adoption and use of open access publishing by academic staff in Nigeria universities. *Journal of Information Science Theory and Practice*, 4(4), 49-63.

Bryan, K. A. & Ozcan, Y. (2020). The impact of open access mandates on innovation. *Review of Economics and Statistics*, In-Press. DOI: https://doi.org/10.1162/rest_a_00926

Chilimo, W., Adem, A., Otieno, A. N. W. & Maina, M. (2018). Adoption of open access publishing by academic researchers in Kenya. *Journal of Scholarly Publishing*, 49(1), 103-122.

Dulle, F. W. & Minishi-Majanja, M. (2011). The suitability of the unified theory of acceptance and use of technology (UTAUT) model in open access adoption studies. *Information Development*, 27(1), 32-45.

Kimbrough, J. L. & Gasaway, L. N. (2016). Publication of Government-Funded Research, open Access, and the Public Interest. *Vanderbilt Journal of Entertainment & Technology Law*, 18(2), 267-302.

Krejcie, R. V. & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurements*, 30, 607-610.

Larivière, V. & Sugimoto, C. R. (2018). Do authors comply with mandates for open access? *Nature*, 562, 483-486.

Singeh, F., Abrizah, A. & Karim, N. (2013). What inhibits authors to self-archive in open access repositories? A Malaysian case. *Information Development*, 29, 24-35. doi: 10.1177/0266666912450450.

Waithaka, M. W. & Onyancha, O. B. (2021). Use of open access channels for scholarly publishing in Kenyan universities. *Publishing Research Quarterly,* In Press. DOI: https://doi.org/10.1007/s12109-021-09795-9

Patterns of Innovation Team of Pharmaceutical in Countries

Chun-Chieh Wang[1,4] , Dar-Zen Chen[2,4] and Mu-Hsuan Huang[3,4]

[1] *wangcc@ntu.edu.tw*
Dept. of Bio-Industry Communication and Development, National Taiwan University,
Taiwan (R.O.C.)

[2] *dzchen@ntu.edu.tw*
Dept. of Mechanical Engineering, National Taiwan University,
Taiwan (R.O.C.)

[3] *mhhuang@ntu.edu.tw*
Dept. of Library and Information Science, National Taiwan University,
Taiwan (R.O.C.)

[4] *Center for Research in Econometric Theory and Applications, National Taiwan University,*
Taiwan (R.O.C.)

Introduction

In the globalized world, external knowledge obtained are facilitated by rising connectivity through organization based and individual based colaboration (Lorenzen & Mudambi, 2013). Individual based knowledge obtained e.g. talent mobility rely on the ability of individuals to overcome geographical distance to exchange and integrate knowledge. Internet have enabled real-time collaboration with skilled located at different location, enhancing the process of leveraging knowledge through direct interactions and personal contacts (Awate & Mudambi, 2018). In this paper, we evaluate the patterns and extent of inventor mobility as the various innovation teams in countries.

Methodology

Patents granted in United States Patent and Trademark Office (USPTO) since 2010 to 2019 and classified as Pharmaceuticals in the World Intellectual Property Organisation technology classification (WIPO, 2009) are collected in this study to analysis patterns of innovation team in countries.

Innovation team in patents are classified as follows:

- Domestic-Team Patent (DTP): All inventors' countries are the same as the assignee's country in a patent.
- Global-Team Patent (GTP): At least one inventor's country is the same as the assignee's country and at least one inventor's country is different from the assignee's country in a patent.
- Foreign-Team Patent (FTP): All inventors' countries are different from the assignee's country in a patent.

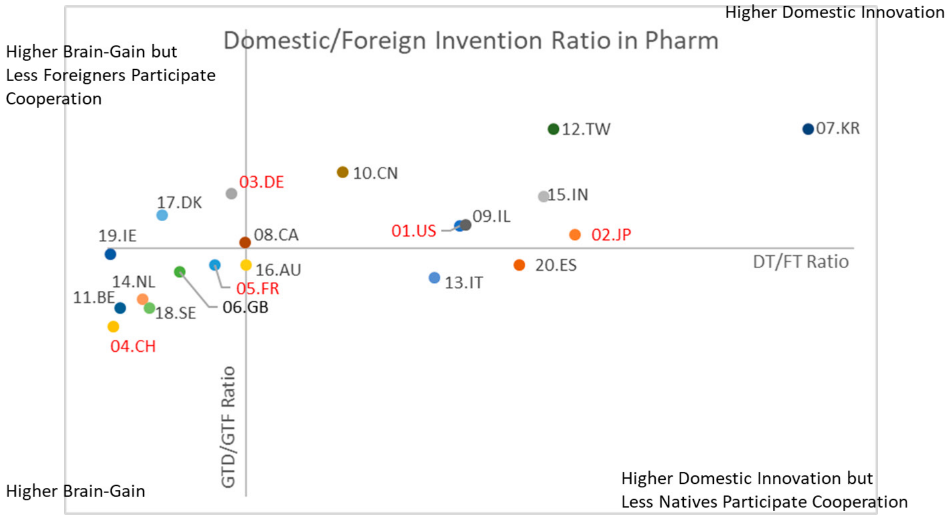
Result

Top twenty patent count of assignee countries with proportions of innovation team patents are listed in Table 1.

Table 1. Innovation-Team Patents of Top 20 Countries in Pharm.

	Patent Count			
	Total	DTP	FTP	GTP
US (UNITED STATES)	40373	84.2%	4.5%	11.3%
JP (JAPAN)	4625	89.5%	3.6%	7.0%
DE (GERMANY)	3519	65.8%	9.9%	24.4%
CH (SWITZERLAND)	2668	17.5%	50.6%	31.9%
FR (FRANCE)	2539	69.2%	12.0%	18.9%
GB (UNITED KINGDOM)	2346	60.9%	15.7%	23.4%
KR (SOUTH KOREA)	1655	93.1%	2.5%	4.5%
CA (CANADA)	1582	61.6%	8.3%	30.1%
IL (ISRAEL)	1217	83.6%	4.4%	12.1%
CN (CHINA)	1216	83.1%	6.6%	10.3%
BE (BELGIUM)	1134	30.7%	44.3%	25.0%
TW (TAIWAN)	1079	79.8%	3.3%	16.9%
IT (ITALY)	1045	78.9%	4.5%	16.7%
NL (NETHERLANDS)	872	52.8%	28.0%	19.3%
IN (INDIA)	863	86.8%	3.7%	9.5%
AU (AUSTRALIA)	826	73.2%	9.8%	16.9%
DK (DENMARK)	819	55.7%	18.9%	25.4%
SE (SWEDEN)	816	55.1%	24.5%	20.3%
IE (IRELAND)	631	13.8%	74.6%	11.6%
ES (SPAIN)	485	81.9%	3.7%	14.4%

*Red: Top 3 Countries in each type of Innovation-Team Patents



*DT: Domestic Team patents; DT: Foreign Team patents; GTD: Domestic inventors in Global Team patents; GTF: Foreign inventors in Global Team patents

Figure 1. Domestic/Foreign Invention Ratio in Pharmaceutical.

Innovation team in pharmaceutical of most countries are DTPs, especially Japan, South Korea, and India get more than 86% pharmaceutical patents all invented by domestic inventors. Ireland, Switzerland, and Belgium are get higher FTP proportion, it means that their pharmaceutical development higher depend on foreign inventors. Switzerland, Canada, and Denmark are with higher proportion of GTP, they more depend on international collaboration in pharmaceutical development.

We posited the pattern of innovation Team in countries as Figure 1, where x axis is the ratio of Domestic Team Patents to Foreign Team Patents and y axis as the ratio of Domestic Inventors of Global Team Patents to Foreign Inventors of Global Team Patents. Countries in the first quadrant are Higher Domestic Innovation; countries in the second quadrant are higher brain-gain but less foreigners participate cooperation; countries in the third quadrant are higher brain-gain; countries in the fourth quadrant are higher domestic innovation but less natives participate cooperation.

Conclusion

There are less study to analysis innovation team in countries according to brain gain and brain drain in inventors. We apply a pilot study to classify the patterns of innovation team of pharmaceutical in countries. In the result we find that Switzerland highly depends on foreign inventors to develop pharmaceutical and also with lower native in international collaboration. South Korea is another extreme country depending on highly domestic inventors in pharmaceutical innovation. The following study will focus on the effect from patents of these various innovation team on native innovation performance, and the findings would be the specific loss of brain drain.

Acknowledgments

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 109L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by the Ministry of Science and Technology (MOST), Taiwan, under Grant No. 109-2634-F-002-045-.

References

Awate, S., & Mudambi, R. (2018). On the geography of emerging industry technological networks: The breadth and depth of patented innovations. *Journal of Economic Geography, 18*(2), 391–419.

Lorenzen,M.,&Mudambi, R. (2013). Clusters, connectivity and catchup: Bollywood and Bangalore in the global economy. *Journal of Economic Geography, 13*(3), 501–534.

WIPO (2009). *IPC and Technology Concordance Table*. Retrieved January 10, 2021 from: https://www.wipo.int/meetings/en/doc_details.jsp?doc_id=117672.

Using Full-text Content of Academic Articles to Classify Research Methods in Library and Information Science

Ruping Wang[1], Liang Tian[2] and Chengzhi Zhang[3]

{[1]wangrp, [2] tianl, [3] zhangcz}@njust.edu.cn

Nanjing University of Science and Technology, Department of Information Management, 200 Xiaolingwei Street, Nanjing 210094 (China)

Introduction

The methodology is a key factor to connect research problems and results. Most of the existing studies have constructed many schemas of research methods in Library and Information Science (LIS). For example, Luo & McKinney (2015) divided the research methodology into the research design, research model, and research theory. Chu & Ke (2017) summarized the research methods of Library Information Science into 16 categories including 'Experiment', 'Bibliometrics', 'Questionnaire', etc.

Identifying the research methods used in academic papers is the basis for analysing usage behaviour and scenarios of different research methods. The research methods of papers usually need to be manually annotated, which is time-consuming and costing a lot. This paper uses a supervised multi-label classification method to identify the research methods in LIS based on the full-text content of academic articles.

Methodology

Dataset

This paper collects research articles from three kinds of journals in LIS field, namely Journal of the Association for Information Science and Technology (JASIST), Library & Information Science Research (LISR), and Journal of Documentation (JOD). After filtering, 5,273 research papers from 1990 to 2019 are obtained for the experimental dataset. Among them, 1,977 papers from 2001 to 2010 were annotated with the research methods by Chu et al. (Chu & Ke, 2017). This annotated corpus is used as training data in this paper. The number of articles using some research methods is very small, leading to data imbalance, and some methods are similar. Hence, we combine 'Bibliometrics' and 'Webometrics' into 'Metrics', and combine the less used research methods as 'Other'. There were finally 8 methods: 'Experiment', 'Theoretical Research Method', 'Questionnaire', 'Content Analysis', 'Metrics', 'Interview', 'Transaction Log Analysis' and 'Other'. Table 1 shows the distribution of different research methods in the training data.

Table 1. The number of articles using different research methods in training set (2001-2010)

No.	Method	Papers' Number
1	Experiment	553
2	Theoretical Research Method	377
3	Questionnaire	366
4	Content Analysis	339
5	Metrics	377
6	Interview	256
7	Transaction Log Analysis	120
8	Other	270

The full-text content of the article is divided into sentences and stop-words are filtered according to a stop-list. Then, the full-text content is represented as a vector after feature selecting by Chi-Square and feature weight computing by TF-IDF.

Classification model of Research methods

Problem transformation and algorithm adaptation (Tsoumakas & Katakis, 2007) are two common kinds of multi-label classification strategies. The strategy of problem transformation includes Binary Relevance, Classifier Chain, Label Powerset, and Random k-label sets. Algorithm adaptation is a way to perform multi-label classification by directly improving or modifying the existing classification algorithm. We use these two multi-label classification strategies to classify the research methods of academic articles in LIS. For problem transformation, we use Naïve Bayes, SVM, and XGBoost as the basic classification algorithms. For algorithm adaptation, we used KNN as the basic classifier. We generate a total of 13 combined classification algorithms, namely BR-NB, BR-SVM, BR-XGB, CC-NB, CC-SVM, CC-XGB, LP-NB, LP-SVM, LP-XGB, RAKEL-NB, RAKEL-SVM, RAKEL-XGB, and MLKNN. We train these combinatorial algorithms respectively and then evaluate each algorithm. We use the optimal algorithm to train the final classifier and then predict the remaining unlabelled articles.

To evaluate and select the optimal classification model, we divide the annotated data into training set, validation set, and test set. The training set accounts for 80% of the total annotated corpus, the validation set and the test set account for 10% respectively. This paper trains each algorithm on the training set and optimize each algorithm on validation set.

In multi-label classification, sample-based evaluation and label-based evaluation are two common evaluation methods(Madjarov et al., 2012). This paper adopts the label-based evaluation method. P, R, and F1 are used as the evaluation indicators. According to the results of the 13 combined classification models, the Classifier Chain based upon XGBoost (CC-XGB) performs the best with the highest F1 value at 80.77%.

Experiments and Results Analysis

Classification result of research methods in LIS

We use CC-XGB to train the final model upon all the annotated articles to predict the remaining 3,296 unlabelled articles. Table 2 shows the prediction results of different methods. We can see that many papers use 'Experiment', 'Theoretical Research Method', 'Questionnaire', 'Content Analysis', 'Metrics' and 'Interview'. 'Transaction Log Analysis' and 'Other' are used less.

Table 2. Distribution of different research methods on all articles (1990-2019)

No.	Method	Papers' Number
1	Experiment	960
2	Theoretical Research Method	1,260
3	Questionnaire	911
4	Content Analysis	734
5	Metrics	1,078
6	Interview	1,080
7	Transaction Log Analysis	167
8	Other	612

Using 1,977 papers that have been manually annotated with research methods and 3,296 papers that have been automatically classified by CC-XGB, we analyse the frequency of different research methods in 1990-1999, 2000-2009, and 2010-2019. The number of papers in each period affects the use of research methods. Hence, we analyse the relative frequency of different research methods in the three periods. Figure 1 shows the result.

We can see that the relative usage frequency of 'Metrics', 'Questionnaire', 'Interview', 'Content analysis' and 'Other' is increasing year by year. The usage of 'Experiment' and 'Transaction Log Analysis' is increasing and then decreasing; the usage of 'Theoretical Research Method' is decreasing gradually. These changes reflect that researchers pay more attention to experiments and data in the process of research.

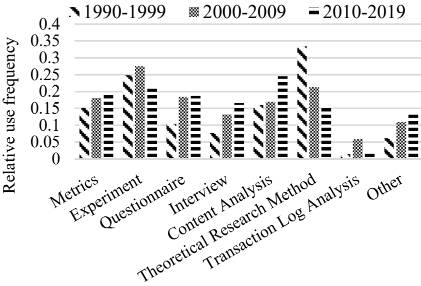


Figure 1. The relative frequency of different research methods in each period.

Conclusion and Future Work

Using LIS field as a case, this paper uses the full-text content of academic articles to classify the research methods. As a preliminary work, we use two multi-label classification strategies to combine with multiple basic classifiers. Finally, the Classifier Chain with XGBoost as the underlying classifier (CC-XGB) has the best performance. The results show that 'Metrics', 'Interview' and 'Theoretical Research Method' occupy a high proportion, and the use of 'Theoretical Research Method' decreased over time.

In the future, we will try to further improve the classification effect. Also, we will make an in-depth analysis of different methods, explore the usage situation and the evolution of different methods.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 72074113).

References

Chu, H. & Ke, Q. (2017). Research methods: What's in the name? *Library & Information Science Research*, 39(4), 284-294.

Luo, L. & McKinney, M. (2015). JAL in the Past Decade: A Comprehensive Analysis of Academic Library Research. *The Journal of Academic Librarianship*, 41(2), 123-129.

Madjarov, G., Kocev, D., Gjorgjevikj, D. & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9), 3084-3104.

Tsoumakas, G. & Katakis, I. (2007). Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining*, 3(3), 1-13.

Algorithm Entities Usage in Chinese Academic Articles from The Domain of Information Science

Yuzhuo Wang[1], Heng Zhang[2] and Chengzhi Zhang[3]

{[1]wangyz, [2] zh_heng, [3]zhangcz}@njust.edu.cn
Department of Information Management, Nanjing University of Science and Technology, Nanjing (China)

Introduction

With the emergence of the fourth paradigm for science, the demand for big data-related algorithms is increasing. As one of the three key elements of artificial intelligence (algorithms, computing power, and data), algorithms have made great contributions to both social science and natural science fields.

An algorithm is any well-defined computational procedure that takes a set of values as input and produces some value as output (Cormen et al., 2009). Academic papers are excellent resources for scholars to learn algorithms. However, manually finding algorithms from massive articles is time-consuming and high-cost. Scholars need automatic methods to extract algorithms. Previous work used rule-based, statistical machine learning, and deep learning methods to extract method entities from academic papers (Hou et al., 2020). However, these studies did not further filter and analyze the extracted entities.

In this article, a deep learning model with a manual filtering method is proposed to find algorithm entities and carry out further exploration. Taking Chinese academic papers in the information science (IS) domain as an example, we plan to explore: *1. How to extract algorithm entities with an automatic method? 2. What is the distribution of algorithms in Chinese IS academic papers?*

We define algorithm entities as nouns or noun phrases representing the name of algorithms or models that are algorithms in nature. Although we only study Chinese IS academic articles, it can be an example to demonstrate that our method is available and can be extended to other disciplines.

Method

We collected the full-text content of papers and then built corpus to train extraction models. Algorithm entities were extracted and reviewed to explore the distribution of algorithm entities in IS papers.

Data collection. The full-text content of 1,367 articles published in the *Journal of the China Society for Scientific and Technical Information* (*JCSSTI*) from 2009 to 2018 was downloaded. *JCSSTI* is the top journal in the field of IS in China. It is believed that articles in this journal are of high quality and algorithms in the articles are representative.

Corpus construction. The papers in PDF format were converted into XML format manually. After that, 200 papers in XML format were selected randomly. Two postgraduates annotated algorithm entities in the 200 papers independently and corrected the inconsistent results. The interrater reliability between the two coders was measured by Cohen's kappa and achieved 0.78. Finally, we got a training corpus that includes 200 articles, 687 algorithm sentences, and 1,130 algorithm entities.

Candidate entities extraction. We chose the commonly used Bi-LSTM+CRF as the extraction model. The training corpus is divided into training and testing set at the ratio of 8:2. After using FastText (https://fasttext.cc/) to train word-vector, training data was inputted to the Bi-LSTM+CRF model. The trained model was then utilized to extract candidate entities from 1,367 articles.

Reviewing algorithm entities. A Ph.D. candidate reviewed all the candidate entities and picked out algorithm entities based on their own experience, authors' descriptions, external knowledge, and experts' explanation. Considering an algorithm could be mentioned with different names, including full name, abbreviation, and aliases, we compiled a dictionary where different names representing the same algorithm were organized into a set.

Analyzing algorithm entities in IS papers.

(1) Classifying algorithms. Algorithms entities were classified into three types: *single, functional and composite*. A s*ingle algorithm* is an explicit algorithm not showing the category and function in the name, such as the *Support vector machine* (SVM). A *functional algorithm* indicates a specific function or type in the name, such as the *Classification algorithm*. A *composite algorithm* contains the actual task solved by the algorithm in the name, such as the *Literature recommendation algorithm based on user interest topics*. In general, A s*ingle algorithm* (SVM) belongs to a f*unctional algorithm* (Classification algorithm), and a *single algorithm* and a *functional algorithm* can belong to the *composite algorithms* (Author name recognition algorithm). The category of the nested algorithm is judged based on the root of the algorithm. For example, the *author name recognition algorithm based on SVM* is marked as *composite* because it is essentially the *author name recognition algorithm.*

(2) Influence of algorithms. The number of papers is the commonly used indicator of influence. We use the same method in Wang's work (Wang & Zhang, 2020) to calculate the influence of the algorithm entity in IS field. For an algorithm *j*,

$$Influence(j) = (\sum_{i=2009}^{2018} \left(\frac{N_{ij}}{N_j}\right))/T_j \qquad (1)$$

Where i represents the year, ranging between 2009 and 2018. N_i is the number of articles published in year i, N_{ij} is the number of articles that mentioned algorithm j in year i. T_j is the duration from the year when algorithm j first appeared in papers to 2018.

Results

Recognition result of algorithm entities

We obtained 5,036 candidate entities from 3,318 sentences in the training corpus, and 30,002 candidate entities from 18,316 sentences in the raw corpus. The performance of the Bi-LSTM+CRF model is precision (67.26%), recall (72.82%), and F1-value (69.93%). After manual filtering, we got the algorithm dictionary containing 2,122 algorithm entities with 4,460 names. All the algorithm entities were classified into 962 *Single algorithms*, 775 *Functional algorithms,* 385 *Composite algorithms*.

Algorithm entities distributed in different year

Figure 1 shows the number (Bar) and proportion (Line) of papers mentioning algorithm entities each year. From 2009 to 2018, about 80% of articles mentioned algorithms. With time going by, the proportion increased from 2009, peaked in 2011, fluctuated from 2011 to 2015, and stabilized at 80% after 2016. It means that algorithms have always been important in the IS academic papers in China.

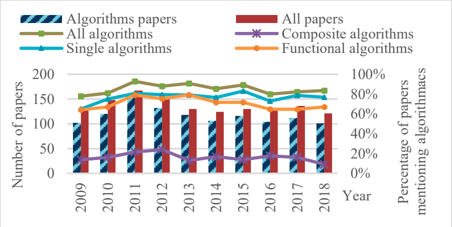


Figure 1. The number and percentage of papers mentioning algorithms

For the three types of algorithm entities, S*ingle algorithms* have always been the algorithm mentioned in the largest number of articles, and its proportion is also the most stable. The percentage of articles mentioning the other two kinds of algorithms decreased significantly after 2012. It indicates that in the IS journal articles, more and more articles begin to directly mention and use the existing single algorithm entities, rather than the descriptive algorithm entities about the purpose and function.

Single algorithms with high influence in IS papers

Considering single algorithms were mentioned most in the articles, we further analyze the influence of every single algorithm based on formula (1). The ten most influential single algorithms are listed in Table 1. Among the ten algorithms, SVM, K-Means and Decision trees are classic machine learning algorithms, the Neural network and Long short-term memory (LSTM) are the deep learning algorithms. It shows that the Chinese IS papers still pay attention to the machine learning task when mentioning algorithm entities. Additionally, the popular algorithms in the computer field also have a strong influence in the IS field in China.

Table 1. Top-10 influential single algorithms

Rank	Algorithm	Rank	Algorithm
1	TF*IDF	6	Neural network
2	Vector space model	7	LSTM
3	SVM	8	K-means
4	Cosine similarity	9	Latent dirichlet allocation
5	Mutual information	10	Decision trees

Conclusion

Taking the algorithms mentioned in Chinese academic articles in IS field as the research object, we propose a method combining deep learning and manual review to extract and evaluate algorithm entities from academic papers. Our research shows that algorithms play an essential role in IS papers in China. The influence of the *single algorithm* is the highest and the most stable. Among the algorithms, classical machine learning algorithms and emerging deep learning algorithms are more influential.

The limitation of this work is we only use a deep-learning model on data from one field. In the future, we will try to combine more deep learning models and improve their performance to reduce the workforce. In addition, the co-occurrence of algorithm entities will be explored to understand the relationship between algorithms. Moreover, the evolution of IS in china can also be analyzed based on the trend of algorithm influence.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 72074113).

References

Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2009). *Introduction to Algorithms, Third Edition*. The MIT Press.

Hou, L., Zhang, J., Wu, O., Yu, T., Wang, Z., Li, Z., Gao, J., Ye, Y. & Yao, R. (2020). Method and Dataset Entity Mining in Scientific Literature: A CNN + Bi-LSTM Model with Self-attention. *arXiv:2010.13583.* http://arxiv.org/abs/2010.13583

Wang, Y. & Zhang, C. (2020). Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. *Journal of Informetrics*, 14(4), 101091.

Breakthrough Foresight Based on Citation Diffusion Features

Haiyun Xu[1], Jos Winnink[2], Xin Zhang[3], Zenghui Yue[4], Jing Li[5] and Chen Liang[6]

[1] *xuhaiyunnemo@gmail.com*
Business School, Shandong University of Technology, Zibo 255000 (China)

[2] *winninkjj@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden 2300 AX (the Netherlands)

{[3] zhangxin@ , [5]lij}@clas.ac.cn
Chengdu Library and Information Center, Chinese Academy of Sciences, Chengdu 610041 (China)

[4] *yzh66123@126.com*
School of Medical Information Engineering, Jining Medical University, Rizhao (China)

[6] *25565853@qq.com*
Institute of Scientific and Technical Information of China (ISTIC), Beijing 100038 (China)

Introduction

The main research questions to be answered in this paper include how can breakthrough innovations be identified using ERTs? Are there different patterns of citation curves between ERTs and their knowledge base? Can the combination of citation curves patterns of ERT and their knowledge base improve the assessing the ERTs' breakthrough potential?

Literature review

Many scholars have conducted studies from different point of views to identify and predict breakthrough innovation. The main commonly used quantitative methods for detecting groundbreaking research frontiers are citation analysis methods (Dahlin & Behrens, 2005; Schoenmakers & Duysters, 2010; Small, 2016), topic mutation analysis(Kleinberg, 2003; Yoon & Kim, 2011), sleeping beauty literature analysis (Van Raan & Winnink, 2018), technical evolution methods and analyses based on machine learning models (Xu & Wang, 2019).

In general, this article aims to identify ERTs that have the potential to become breakthrough innovations with the method of citation curve fitting analysis and the characteristics of citation diffusion. By extracting the patterns and characteristics of citation curves of different ERTs, we expect to supplement and enrich the identification and prediction methods of breakthrough innovation. The breakthrough innovation mentioned in this paper refers to the technological innovation with the characteristics of dynamic, discontinuity and novelty.

Methodology

Theoretical base

This paper aims to evaluate the breakthrough potential of ERTs through analyzing the features of development trend and the correlation between the citation curve of ERTs and their knowledge base. Based on the analysis, we can finally achieve the breakthrough prediction. According to this purpose, we first list four theory bases of using citation curves to track the development patterns of ERTs, and discusses the feasibility of using citation curves to achieve breakthrough prediction from these four theory bases.

Procedural steps

There are three main steps in this article to achieve breakthrough innovation predictions for ERTs. First, we identify ERTs in a field, and then perform fitting analysis on the citation curves of these topics and their knowledge base. Finally, through comparative analysis of different citation curve patterns, we assess the breakthrough potential of different ERTs. The following is the analysis process in this article.
The content of method model in this paper includes identifying features and connotations, as well as the measurement characteristics and methods (Table 1).

Table 1. Method models for breakthrough identification of ERTs

Identify features	Measurement characteristics
ERTs	High novelty and rapid growth
basic knowledge	The knowledge base continues to develop
Knowledge transition	Multiple layers of citation curves
Continuous growth	Keep high citations and continue growing in recent period, the highest citation peak appears late

Empirical analysis

Data acquisition and analysis

The study of stem cells—a type of biological cell capable of self-renewal and multidirectional differentiation—is an important topic in biomedical research. The literature on stem cells research was selected as testbed to demonstrate the approach proposed in this paper. The data for this study were collected on October 20, 2018.

Comparison of Citation Curve Features

We then constructed the citation curves of the knowledge base in the form of time slices from 2001 to 2017 as the "citation curve of knowledge base". The citation curve of the ERTs in this study is from 2001 to 2017, but the publication of the knowledge base is from 1990 to 2015. Therefore, there are two piece of curve differences, which will not affect the analysis. The knowledge base crosses a relatively long time-span, to better understand and comparatively analyze the development pattern of the knowledge base and ERTs, we divide the knowledge base into EKB (1990-2000) and late knowledge base (LKB) (2001-2015).

Comparison of development pattern of knowledge base and ERTs in different time periods

Taking topic 2276 as an example (Figure 1), its citation curves of the EKB show a fluctuating trend. The curve reaches its local peaks in 2005, 2010, 2013, and 2015 respectively, and their peak value are not much different (Figure 1). The citation curve of the LKB has roughly experienced three transitions (leaping), in 2011, 2013 and 2015 respectively, and the highest peak of the citation curve appeared in 2017 in the 2015 papers' curve. The recent citation curve shows an uprising trend, and the citations value is higher than before. There is a possibility of another curve transitions. These characters indicate that the LKB of the ERT is in developing. The citation curve of ERTs has roughly undergone three transitions, in 2006, 2011 and 2013 respectively, and the highest peak of the citation curve appeared in 2017 in the 2015 papers' curve. The recent citation curve shows an uprising trend, and the popularity is higher than before. These characters indicate that this ERT is in development.

Discussion and conclusions

Based on the citation diffusion characteristics, combining the citation curve analysis of both knowledge base and ERTs, we can better analyze the development model of ERTs and assess their breakthrough potential, to provide reference and support for making science and technology planning. Therefore, the use of citation curves to identify breakthroughs in ERTs can further improve the theory and methodology of breakthrough identification. The professional evaluation of empirical results by experts and the current research status in the stem cell field prove that citation curve analysis combining knowledge base and ERTs can more efficiently assess the breakthrough potential of ERTs.

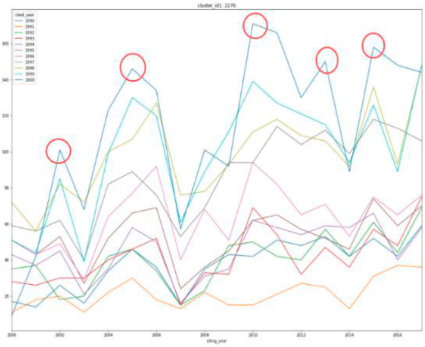


Figure 1. Citation curve of the EKB (# 2276)

Acknowledgments

This article is an outcome of the projects (No. 71704170, No. 71704063), both supported by the National Natural Science Foundation of China, and Tai Shan Scholar Foundation in Shandong province (202103069).

References

Dahlin, K. B. & Behrens, D. M. (2005). When is an invention really radical?: Defining and measuring technological radicalness. *Research Policy*, 34(5), 717-737.

Schoenmakers, W. & Duysters, G. (2010). The technological origins of radical inventions. *Research Policy*, 39(8), 1051-1059.

Small, H. (2016). Referencing as cooperation or competition. Theories of Informetrics and Scholarly Communication: A Festschrift in Honor of Blaise Cronin, 49-71.

van Raan, A. F. J. & Winnink, J. J. (2018). Do younger Sleeping Beauties prefer a technological prince? *Scientometrics*, 114(2, SI), 701-717.

Xu, L. & Wang, F. (2019). Scientific Frontier Prediction Model Based on Support Vector Machine and Improved Particle Swarm Optimization. *Information Sciences*, 37(8), 22-28.

Yoon, J. & Kim, K. (2011). Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. *Scientometrics*, 88(1), 213-228.

Analysis of selected journals in Excellence Action Plan for China STM Journals Based on Bibliometrics

Yang Rui[1] and Wang Baoji[2]

[1] *664681261@qq.com*

[2] *wbj@cau.edu.cn*
Information Research Center, CAU Library, China Agricultural University, Beijing (China)

Introduction

Excellence Action Plan for China STM Journals was launched in 2019. The project is by far the largest project in China which aims to promote the high-quality development of Chinese scientific journals. Compared with the previous scientific journals funded projects in China, Excellence Action Plan for China STM Journals has the following new characteristics:

1). The number of the funded journals is the largest in a single funding cycle with 280 journals.

2). The amount of funding is the largest in a single funding cycle. The total funding is expected to exceed 160 million dollars.

3). The project is actually a systematic integration and reconstruction of the previous scientific journals funded projects.

4). The project has dynamic management and relatively more scientific evaluation.

Excellence Action Plan for China STM Journals has 7 sub-projects, including 4 journal sub-projects, 1 sub-project of the pilot for clustering journals, 1 sub-project of the international digital publishing service platform, 1 sub-project of selecting and breeding high-level talents for running journals. We focus on the analysis of 4 journal sub-projects.

The 4 journal sub-projects include leading journals (22), key journals (29), echelon journals (199) and new high start journals (30). Leading journals, key journals and new high start journals are all English journals. 199 echelon journals include 99 English journals, 68 Chinese basic research journals, 27 Chinese engineering and technology journals, and 5 Chinese popular science journals. These journal sub-projects are arranged like a pyramid. Specifically, leading journals are at the top of the pyramid which are the most promising to become world-class journals by some measures and they were selected by subject areas; key journals and leading journals form a competitive situation; the project positions echelon journals as a backup force; new high start journals should found new English scientific journals of the traditional predominant, emerging multi-disciplinary, strategic frontier or key generic technologies fields.

At present, the evaluation of journals in different languages is an urgent problem to be solved, especially for non-native English speaking countries. Dahler-Larsen proposes the PLOTE-index which could measure the percentage of citations flowing from the non-English publications of the researchers. This study analyzes the bibliometric characteristics of the selected journals based on new indicators being explored in China.

Method

The analysis of the academic impact of the selected journals is mainly based on *Annual Report for Chinese Academic Journal Impact Factors 2019* and *Annual Report for World Academic Journal Impact Factors 2019* which were both issued by China National Knowledge Infrastructure (CNKI).

Annual Report for Chinese Academic Journal Impact Factors 2019 proposes the Academic Journal Clout Index (CI), a comprehensive evaluation indicator that reflects the academic impact of journals. This indicator combines total cites and impact factors, and the Journal Mass Index (JMI) is used to modify this indicator. *Annual Report for Chinese Academic Journal Impact Factors 2019* quarters the Chinese journals of each discipline in descending order of CI.

Annual Report for World Academic Journal Impact Factors 2019 proposes World Academic Journal Clout Index (WAJCI) which could be got by dividing the CI by the median CI. *Annual Report for World Academic Journal Impact Factors 2019* also quarters the journals in the world in descending order of WAJCI. Multi-disciplinary journals are counted according to its highest quartile.

Results

Table 1 shows the international and domestic academic impact of the selected journals. According to *Annual Report for Chinese Academic Journal Impact Factors 2019*, 14 leading journals are ranked Q1, which means that 63.64% of leading journals have high academic impact in China.

According to *Annual Report for World Academic Journal Impact Factors 2019*, 15 journals are ranked Q1, which means that 68.18% of leading journals have high international academic impact. It is found that high international academic impact is not a necessary condition for selected, and the selection of leading journals is more focused on English journals with high international academic impact.

Table 1. The international and domestic academic impact of the selected journals.

Report	Quartile	Leading journals	Key journals	Echelon journals			
				Summary	English journals	Chinese basic research journals	Chinese engineering and technology journals
Annual Report for Chinese Academic Journal Impact Factors 2019	Q1	63.64% （14/22）	24.14% （7/29）	72.36% （144/199）	53.54% （53/99）	95.59% （65/68）	96.30% （26/27）
	Q2	13.64% （3/22）	27.59% （8/29）	9.05% （18/199）	14.14% （14/99）	4.41% （3/68）	3.70% （1/27）
	Q3	9.10% （2/22）	24.14% （7/29）	4.52% （9/199）	9.10% （9/99）	-	-
	Q4	-	10.34% （3/29）	4.02% （8/199）	8.08% （8/99）	-	-
	No data available	13.64% （3/22）	13.79% （4/29）	10.05% （20/199）	15.15% （15/99）	-	-
Annual Report for World Academic Journal Impact Factors 2019	Q1	68.18% （15/22）	41.38% （12/29）	16.08% （32/199）	12.12% （12/99）	20.59% （14/68）	22.22% （6/27）
	Q2	31.82% （7/22）	51.72% （15/29）	33.67% （67/199）	40.40% （40/99）	33.82% （23/68）	14.81% （4/27）
	No data available	-	6.90% （2/29）	50.25% （100/199）	47.47% （47/99）	45.59% （31/68）	62.96% （17/27）

Note: Five Chinese popular science journals are not academic journals, so these journals are not listed in Table 1.

According to *Annual Report for Chinese Academic Journal Impact Factors 2019*, 7 key journals are ranked Q1, which means that 24.14% of key journals have high academic impact in China. Under the same calculation conditions, the domestic academic impact of key journals is lower than that of leading journals.

According to *Annual Report for World Academic Journal Impact Factors 2019*, of the 29 selected key journals, 27 journals have available data which are all ranked Q1 or Q2. Whereas, under the same calculation conditions, the international academic impact of key journals is lower than that of leading journals. And the selection of key journals is more focused on English journals with higher international academic impact.

According to *Annual Report for Chinese Academic Journal Impact Factors 2019*, 144 of the 199 echelon journals are ranked Q1, which means that 72.36% of the journals have high domestic academic impact. From the overall data, the domestic academic impact of echelon journals is higher than that of leading and key journals, mainly because half of the echelon journals are Chinese journals, and the proportion of these Chinese journals ranked Q1 is as high as 95%.

According to *Annual Report for World Academic Journal Impact Factors 2019*, 32 of the 199 echelon journals are ranked Q1, and 16.08% of the journals have high international academic impact. From the overall data, the international academic impact of echelon journals is much lower than that of leading and key journals, and there is a big difference between the international academic impact of echelon journals and that of world-class journals. However, the proportion of English journals of echelon journals ranked Q1 is 12.12%, which is lower than that of Chinese journals (20%), indicating that some Chinese journals have already accumulated definite international academic impact.

Conclusion and Discussion

Leading journals and key journals are selected by priority construction areas; are more focused on English journals with high international academic impact. Echelon journals give consideration to both Chinese and English scientific journals; do not have overall high international academic impact, while the Chinese journals among them have high domestic academic impact.

So far, the project has brought many positive effects, the most important one is that it points out the macro planning for the development of Chinese scientific journals. However, we should also note the limitations of this project. For example, although 280 journals have been funded, the coverage is still very limited in terms of the number of scientific journals in China as a whole.

References

Peter, D. L. (2018). Making citations of publications in languages other than English visible: On the feasibility of a PLOTE-index[J]. *Research Evaluation,* 3, 3.

China Scientometrics and Bibliometrics Research Center, Tsinghua University Library (2019). Annual Report for World Academic Journal Impact Factors 2019. China Academic Journals Electronic Publishing Corporation.

China Scientometrics and Bibliometrics Research Center, Tsinghua University Library (2019). Annual Report for Chinese Academic Journal Impact Factors 2019. China Academic Journals Electronic Publishing Corporation.

Does knowledge of Library and information science field transform to technology?

Xiao Yang [1], Lingzi Feng [2] and Junpeng Yuan [3]

{[1] yangxiao, [2] fenglingzi , [3] yuanjp}@mail.las.ac.cn
National Science Library, Chinese Academy of Sciences (China)
Department of Library, Information and Archives Management, School of Economics and Management,
University of Chinese Academy of Sciences (China)

Introduction

Scientific research is the source and driving force of technological innovation. (Bush, 1945). The transformation from science to technology is more common in hard disciplines (Lo, 2010; Sung et al., 2015), while soft disciplines are not the main source of knowledge for technological innovation. Some scholars have investigated the impact of citing humanities and Social Sciences papers on patents (de Moya-Anegon et al., 2020), and demonstrate the technological and economical viability of Library Science (Halevi, 2012).Therefore, this paper studies the relationship between the papers in LIS and the its citing patents, in hopes of understanding the characteristics of the knowledge transformed from LIS to technology, so as to explore the new advantages of the development of LIS from the perspective of technological innovation.

Data

This study focuses on the papers in the field LIS. Firstly, we selected the Web of Science (WoS) platform and "Information Science & Library Science" in the JCR classification system, which included 87 journals. Then, we matched 417 papers that are cited by patents in The Lens (covering 22 journals in LIS), and found 1,437 patents cited these papers. In order to better reveal the situation of these LIS papers cited by patents, we choose the subset of papers included in WoS as the analysis object. By using the paper identifier such as DOI to match the papers in WoS, 357 papers to be analyzed are matched. Finally, we matched the patents which cited these 357 papers in The Lens again, and 1090 patents were obtained.
All data were acquired on December 11st, 2020.

Analysis of the linkages between scientific publications and patents in the field of LIS

Journal distribution

The journal distribution of LIS papers cited by patents are shown in Table 1. These 357 papers cover 22 journals (25.3%) in this field, among which Information & Management has the most number of papers cited by patents. Through the Spearman correlation test, it is found that the correlation coefficient is 0.657 and the significance is 0.01, indicating a significant positive correlation. That is, the higher the impact factor of the journal, the higher the number of the patent citation.

Table 1. The journal distribution of LIS papers cited by patents (top10)

Journal	*Num. of articles*	*Cited by Patent*	*Journal Impact Factor*
Telecommunications Policy	28	246	1.7638
Information & Management	60	151	3.729
Electronic Library	20	109	0.6796
Information Systems Research	37	86	2.8306
Online Information Review	26	75	1.6188
Scientometrics	36	64	2.4082
Social Science Computer Review	17	61	2.5378
Library Hi Tech	9	35	0.958
Library & Information Science Research	8	17	1.3394
Information Society	5	14	1.8036

Table 2. Annual Time lag between LIS paper and its citing patent. (recent ten years)

Pub. year (Patent)	*Time lag*		
	Max.	*Min.*	*Average*
Average	18.5	3	9.8
2020	32	3	15.8
2019	33	0	14.8
2018	32	4	16.0
2017	28	2	15.6
2016	27	1	14.5
2015	27	3	13.5
2014	29	2	13.7
2013	34	0	12.6
2012	32	1	12.8
2011	27	2	13.8
2010	25	2	10.5

Analysis on time lag of patent-paper

In order to accurately calculate the time lag of LIS papers cited by patents and its citing patent, we count the time lag between the publication year of each patent and the publication year of one or more LIS papers cited by this patent and observed the time lag between patents and LIS papers every year. As shown in Table 2, the average time lag is 9.8 years. According to the average annual time lag of this study, the time lag has become longer and longer from 1989 to now. By 2020, the average annual time lag has reached 15.8 years, which is 6 years longer than the overall average time lag.

Analysis on the topic flow from paper to patent

In order to sort out the flow of scientific knowledge in LIS to technology, we used LDA model to determine the topic of the LIS paper, including: Library and public services(1), literature citation and author cooperation(2), decision support and user communication(3), network service and reference consultation(4), website production and web design(5), knowledge organization and strategic planning(6), project management and organization(7), financial marketing and economic mechanism(8) and others(0). According to the International Patent Classification (IPC) the topics(categories) of patent are: physics(G), electricity(H), human life needs(A), operation and transportation(B), and chemical metallurgy(C). Then we used Sankey diagram to present the paper-patent topic flow (Figure 1).

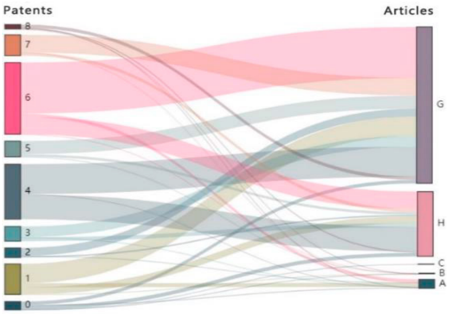


Figure 1. Sankey diagram of paper-patent topic flow.

Quality relationship of paper-patent

In order to further reveal the correlation between LIS papers' impact and its citing patents' quality, we also present this relationship through Sankey diagram (Figure 2). The column on the left side is the cited frequency of papers (cited by papers), and the column on the right side is the cited frequency of patents (cited by patents). Since the data of cited frequency is relatively dense, we divided the data into few sections based on the size of number. As can be seen from Figure 2, no matter the impact of the LIS papers is big or small, they are cited by different quality patents.

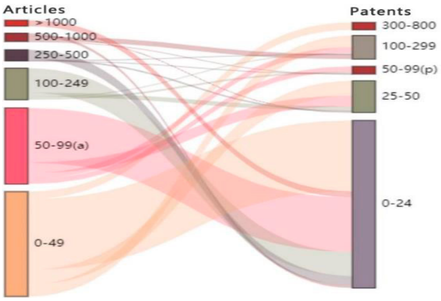


Figure 2. Sankey diagram of paper-patent quality flow.

Conclusion

As a comprehensive interdisciplinary soft discipline, the Library and information science (LIS) does cited by patents, and we could explore new growth points by analyzing the situation and characteristics of the conversion of scientific knowledge in this field to technology.

Acknowledgments

Thank Ruhao Zhang for his support in data analysis. This work is supported by LIS Outstanding Talents Introducing Program, Bureau of Development and Planning of CAS [2018] No. 12.

References

de Moya-Anegon, F., Lopez-Illescas, C., Guerrero-Bote, V. & Moed, H. F. (2020). The citation impact of social sciences and humanities upon patentable technology. *Scientometrics,* 125(2), 1665-1687. doi:10.1007/s11192-020-03530-5

Halevi, G. M., Moed, H. F. (2012). Patenting library science research assets. *Research Trends,* 27.

Lo, S.-c. S. (2010). Scientific linkage of science research and technology development: a case of genetic engineering research. *Scientometrics,* 82(1), 109-120. doi:10.1007/s11192-009-0036-8

Sung, H.-Y., Wang, C.-C., Huang, M.-H. & Chen, D.-Z. (2015). Measuring science-based science linkage and non-science-based linkage of patents through non-patent references. *Journal of Informetrics,* 9(3), 488-498. doi:10.1016/j.joi.2015.04.004

Bush,V. (1945). *Science and the Endless Frontier*: Washington, DC: national foundation.

Artificial Intelligence in Pharmaceuticals: Bibliometric and Collaboration Network Analysis of Patents

Yang Xin

yang_xin@fudan.edu.cn
Fudan University, Fudan University Library, Intellectual Property Information Service Centre, No. 300 Guonian Road, 200433 Shanghai (China)

Introduction

Artificial Intelligence (AI) includes four categories: systems that think humanly (the cognitive modeling approach), systems that act humanly (the Turing Test approach); systems that think rationally (the "laws of thought" approach), and systems that act rationally (the rational agent approach) (Russell & Norvig, 2010). With the thriving development of AI techniques, more than 340,000 AI-related inventions have been filed applications up to early 2018 (WIPO, 2019). Through the bibliometric analysis of patent literature, it can objectively and quantitatively investigate the stage of technical improvement, predict the technological and economic development trend. Thus, both bibliometric and Social Network Analysis (SNA) are conducted to achieve a holistic view of the AI-pharmaceutical field. The study is intended to provide references for future strategy planning, development, and technological marketization.

Method and Data

The patent data is retrieved from Derwent Innovation Index database (DII). The patent retrieval strategy is composed of two portions: (1) search terms related to AI (Topic "artificial intelligen*" or "depth learning*" or "deep learning*" or "natural language processing*" or "speech recognition*" or "computer vision*" or "gesture control*" or "smart robot*" OR "Video recognition*" OR "Voice translation*" OR "Image Recognition*" OR "machine intelligen*" or "Machine learning*"); (2) Derwent Manual Codes related to pharmaceuticals (B01* to B15*). The date of retrieval is December 16, 2019. As a result, 1282 raw records are achieved and downloaded from DII. Then, the patent family records are displayed as a single patent, and duplicates are manually cleaned. Finally, 3610 items of single patent information are obtained.

Results

Patent activity trend

The application year of patents can reflect inventive performance. Figure 1 exhibits the annual patent application count in the AI-pharmaceutical domain. It is observed that the number curve of patent applications could be divided into two stages, which presents diverse growth patterns. The first phase is from 1972 to 2010, when the number of patent applications had endured four decades of slow development stages, with the annual quantity less than 100. The second phase is during the interval from 2011 to 2017, when the count of patent applications has shown a sharp increase, with an average growth rate of 42.5%. This intense rise is attributed to the dramatic promotion of artificial intelligence, in particular, deep learning sprang up as an emerging field of machine learning, which tremendously boosts the development in visual object recognition, drug discovery, and many other domains (LeCun, Bengio & Hinton, 2015). After 2017, the decline in quantity is caused by the 18 months lag period of patent applications. Therefore, the number of AI-pharmaceutical-related patents in both 2018 and 2019 is inadequate, for reference only.

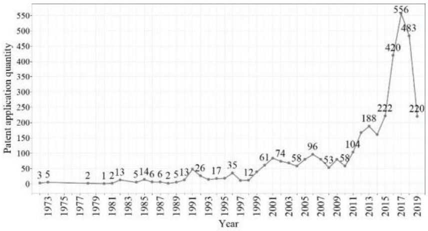


Figure 1. Annual amount of AI- pharmaceutical-related patent applications

Geographical distribution

The information of patent geographical coverage can provide an insight into the industry activity and market competition. The main geographical layout of AI-pharmaceutical-related patents are displayed in Figure 2. The earliest priority country refers to the country or region where the patent application is first filed, suggesting the origin of inventions and national innovation activity. The target country stands for any country or region where the patent right is subsequently protected, indicating the competitive market.

The countries or regions of first fillings are the United States of America (US) with 2248 number of records, followed by China (CN, 364 records), Japan (JP, 233 records), European (EP, 152 records), and Republic of Korea (KR, 116 records). The patent quantity of the US is more than 6 times larger than the second productive country (China), reflecting its technology predominance in the AI-pharmaceutical field. For the top 10 earliest priority countries or regions, apart from the developing country China (rank 2nd), the rest are all developed countries or regions.

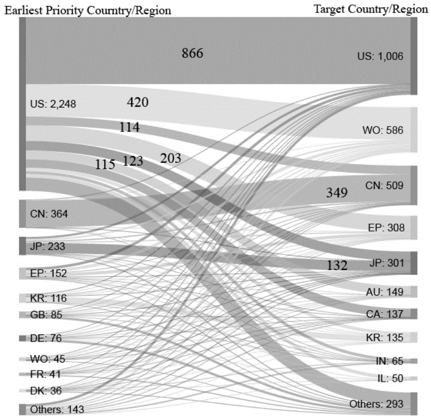


Figure 2. Main geographical distribution of the AI-pharmaceutical-related patents

Patents cooperation network

The collaboration networks, established by assignees of co-owner patents, have been proven to evaluate what positions organizations occupy in a special technology field (Li, Garces & Daim, 2019). The collaborative network is composed of 1233 nodes and 450 links (Figure 3).

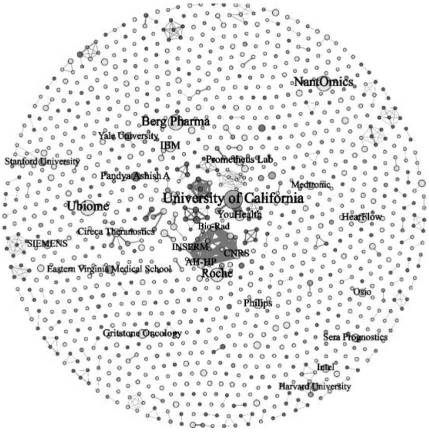


Figure 3. Assignees cooperative network of AI-pharmaceutical patents

The average density of the network is merely 0.001. The low density of collaborative network suggests that there is a weakly cooperative relation among assignees, due to plenty of unconnected nodes (67% of all assignees). Approximately 12% of patents are co-owned by assignees from at least two organizations. This also implies that the co-owner relationships of AI-pharmaceutical-related patents are not highly cooperative.

The patents engaged by companies account for around 73 percent of the total. Companies and academic institutions are the friskiest assignees in the applications of AI-pharmaceutical-related patents. The co-owned patents mostly exist between enterprise and academic institutions, while the co-ownership patents among enterprises are marginal. The cooperative relationships between enterprise and enterprise are basically concentrated upon inter-group and parent-subsidiary.

Conclusions

The quantity of AI-pharmaceutical-related patents has entered a sharp growth period since 2011. The United State of America is both the dominant nation of technological innovation and the significant AI-pharmaceutical market, followed by China. Additionally, the US pays higher attention to overseas patent layouts than other countries or regions. The low-density value of the assignee cooperation network indicates the insufficiency of inter-organizational collaboration. In the AI-pharmaceutical field, companies and academic institutions are the friskiest innovation subjects. The academic-enterprise collaboration should be enhanced to accelerate technology transform, and it's beneficial to consider the geographical factor.

Acknowledgments

This study was funded by “CALIS National Information Center in Medicine” (No. CALIS-2020-02-007).

References

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Li, S., Garces, E. & Daim, T. (2019). Technology forecasting by analogy-based on social network analysis: The case of autonomous vehicles. *Technological Forecasting and Social Change*, 148, 119731.

Russell, S. & Norvig, P. (2010). *Artificial Intelligence A Modern Approach* (3rd). London: Pearson Education.

WIPO. (2019). *WIPO Technology Trends 2019: Artificial Intelligence*. Retrieved July 2, 2020 from: https://www.wipo.int/tech_trends/en/artificial_intelligence/.

Who Shares Scholarly Output on Facebook?

Houqiang Yu[1], Wei Zhang[2], Yue Wang[3] and Tingting Xiao[4]

[1] yuhouq@yeah.net, [2] wei961014@163.com, [3]595905250@qq.com
School of Economics & Management, Nanjing University of Science and Technology, Nanjing (China)

[4]601604391@qq.com
Nanjing library, Nanjing (China)

Introduction

Scholarly outputs are increasingly shared on Facebook (Bowman, 2015). Therefore, Facebook becomes one important altmetrics data source. We use *scientific Facebook user* to refer to Facebook user who mention scholarly output in their posts. This study has revealed scientific Facebook users' productivity distribution based on a large dataset. Moreover, the study has selected different samples of scientific Facebook users and investigated their identity distribution. The research questions are:(1) What is the productivity distribution of scientific Facebook users? (2) What are the identities of scientific Facebook users?

Data and Method

Data are collected from Altmetric database. Facebook mentions of scholarly output from September 1st, 2017 to August, 31st, 2020 were retrieved. In total, there were 1,506,036 Facebook posts with 1,613,993 mentions of scholarly outputs.

The study used simple random sampling and stratified sampling strategy to extract the data. In total, five samples were obtained, consisting of 2,160 records, including 500 random records (sample 1) and 1,660 stratified records (sample2 to sample 5). Qualitative coding analysis was adopted on these data to reveal the identity of scientific Facebook users. As shown in Figure 1, there are six basic types of identities. These basic types are further divided into 30 specific types of identities. While the basic types of identities are obtained in a top-down way, the specific types of identities are summarized in a bottom-up manner.

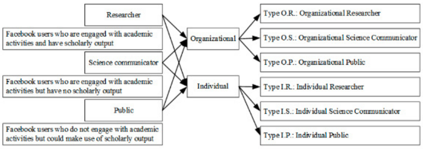

Researcher
Facebook users who are engaged with academic activities and have scholarly output
Science communicator
Facebook users who are engaged with academic activities but have no scholarly output
Public
Facebook users who do not engage with academic activities but could make use of scholarly output
Organizational
Individual
Type O.R.: Organizational Researcher
Type O.S.: Organizational Science Communicator
Type O.P.: Organizational Public
Type I.R.: Individual Researcher
Type I.S.: Individual Science Communicator
Type I.P.: Individual Public

Figure 1. Six basic types of scientific Facebook user

Results

Productivity distribution of scientific Facebook users

Number of mentioned scholarly outputs is used to measure scientific Facebook user's productivity. As shown in Table 1, scientific Facebook users are further divided into 20 levels of activeness according to their productivity with each level having around 5% of total scientific Facebook posts.

Table 1. Scientific Facebook users of different levels of activeness

A.	*Range*	*P.U.*	*A.*	*Range*	*P.U.*
A1	[1,8]	67.9%	A11	(222,294]	0.5%
A2	(8,18]	11.1%	A12	(294,387]	0.4%
A3	(18,29]	6.0%	A13	(387,488]	0.3%
A4	(29,43]	4.0%	A14	(488,629]	0.2%
A5	(43,60]	2.7%	A15	(629,858]	0.2%
A6	(60,82]	2.0%	A16	(858,1310]	0.1%
A7	(82,108]	1.4%	A17	(1310,1981]	0.1%
A8	(108,139]	1.2%	A18	(1981,3288]	0.1%
A9	(139,175]	0.9%	A19	(3288,5648]	0.04%
A1	(175,222]	0.7%	A20	(5648,14466]	0.01%

**A.* is activeness of scientific Facebook users, *P.U.* is percentage of users.

Identities distribution based on the random sampling strategy

Based on the random sampling strategy, the coding result is shown in Figure 2. Organizational users are dominant in scientific Facebook users, taking up 82.5%. Public users (type O.P. and I.P., 54.7%) are the majority, followed by researcher users (type O.R. and I.R., 32.1%) and science communicator users (type O.S. and I.S., 13.2%). To be more specific, organizational public users (type O.P.) plays a leading role in scientific Facebook users.

Zoom into Figure 2, Figure 3 is obtained. Among organizational public users (type O.P.), business organization has taken up 25.7%. It's followed by social organization (18.5%) and government organization (2.3%). Among organizational researcher users (type O.R.), higher education institutions (O.R.1) and academic associations (O.R.2) are two major types, taking up 10% and 7% respectively. Among organizational science communicator users (type O.S.), scholarly journals (O.S.3) and scientific websites (type O.S.6) have relatively higher percentage, that is 3.4% and 2.8% respectively. For individual users, university faculties (I.R.1) and non-academic personal media (I.P.2) are leading types of user, taking up 5.1% and 4.7% respectively.

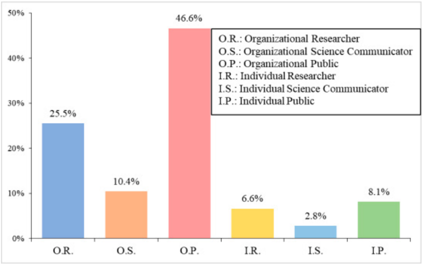


Figure 2. Identities distribution of scientific Facebook users based on general sample

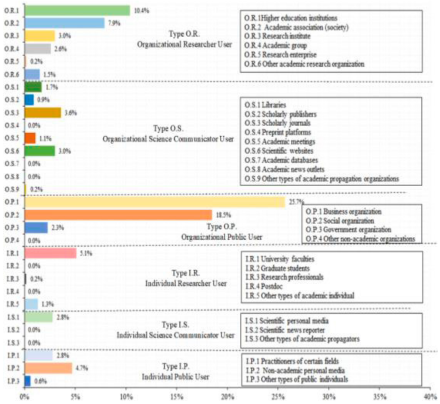


Figure 3. Specific identities distribution of each basic user type

Identities distribution based on the stratified sampling strategy

Identities distribution of different stratified sample users has presented different pattern, as shown in Table 2. For example, type O.S. percentage is decreasing rapidly from sample 2 to sample 5. In sample 2, the percentage of type O.S. user is astonishingly as high as 63%, but in sample 5, the percentage is as low as 8.8%. It suggests that the productivity of scientific Facebook users will have great impact on their identities distribution.

Table 2. Identities distribution of users across samples of different level of activeness

Code	***Sample 2 (A16-A20)***	***Sample 3 (A11-A15)***	***Sample 4 (A6-A10)***	***Sample 5 (A1-A5)***
O.R.	16.1%	26.0%	27.5%	23.1%
O.S.	63.2%	49.1%	27.0%	8.8%
O.P.	1.9%	10.2%	25.5%	49.5%
I.R.	9.7%	5.5%	6.8%	7.5%
I.S.	8.4%	7.9%	8.3%	2.9%
I.P.	0.6%	1.3%	5.0%	8.4%

Conclusion

Productivity distribution is highly skewed towards the lowly productive scientific Facebook users. Identities of scientific Facebook users are very diversified in types. They can be classified into six basic types and 30 specific types. Organizational scientific Facebook users are prevailing taking up more than 80% of total users. Public users (Ptg= 55%) have surpassed researcher users (Ptg= 32%) and are the major type of scientific Facebook users.

Scientific Facebook user's level of activeness has strong connection with their identities. There is clear increasing pattern as regards percentage of public users with the decreasing level of activeness. These findings are based on larger sample and more specific codings compared with previous studies (Mohammadi, Barahmand & Thelwall, 2020). By not considering the filed factor in sampling strategy, the study has revealed the overall situation of Facebook users, but it can also be regarded as a major limitation because disciplinary difference exists.

Acknowledgement

The research is supported by the National Natural Science Foundation of China (NO. 71804067), the Youth Funding of Social Science of Ministry of Education (18YJC870023), the Fundamental Research Funds for the Central Universities of China (No. 30920021203), Postgraduate Practice Innovation Project of Jiangsu Province (SJCX20_0144). The authors would like to thank Altmetric.com for providing the data.

References

Bowman, T. D. (2015). *Investigating the use of affordances and framing techniques by scholars to manage personal and professional impressions on twitter*. Bloomington, IN: Indiana University.

Mohammadi, E., Barahmand, N. and Thelwall, M. (2020), Who shares health and medical scholarly articles on Facebook?. *Learned Publishing*, 33, 111-118.

Where, how and why are scientific products mentioned in policy documents?

Houqiang Yu[1], Biegzat Murat[2], Jiatong Li[3] and Longfei Li[4]

[1] yuhouq@yeah.net, {[2] nulibiegzat, [4] nust_llf}@163.com, [3]1065047683@qq.com
School of Ecnomics & Management, Nanjing University of Science & Technology, Nanjing (China)

Introduction

Policy document mention is considered to be useful for measuring the societal impact of scholarly output. Major altmetric databases such as Altmetric and PlumX have collected policy document mentions for several years. Policy document is defined in a broad sense that a range of document types such as government guidelines, reports, white papers, recommendations and independent policies are all included. In this study, publications that are unrelated to policy (such as doctoral thesis, conference proceedings and conference reports, etc.), or just marginal part of the policy document (such as brief, appendices and lists, etc.) are not considered as policy documents and will not be further analyzed.

Previous studies have investigated the coverage (Bornmann, Haunschild, & Marx, 2016) and data quality (Yu et al., 2020) of policy document mentions, as well as the predicting factors of whether scientific products get mentioned by policy document (Kale et al., 2017). However, it is not clear how policy document mentions can be interpreted. This study is aimed to answer three research questions. (1) Where are scientific products mentioned in the policy documents? (2) How are scientific products mentioned? (3) Why are scientific products mentioned?

Methodology

Policy document mention data were retrieved from Altmetric database on August 19th, 2020. In total, there were 2.22 million mentions from 187 institutions. Considering the difference between institutions, we first randomly selected 40 institutions (as shown in Table 1) and then randomly selected 10 records from each institution. The sample dataset consists of 400 records.

The analysis was conducted with the following steps. (1) Click the URL of policy document in the record and download the file. (2) Search for the mentioned scientific product. Title, part of the title, source, author's first name, DOI, publication link, or publication date was used to help locate the mentioned scientific product. (3) Check in what format the policy document has mentioned the scientific product. (4) Examine in which part the mention appears in the policy document. (5) Investigate what part of the scientific product is mentioned. (6) Determine the motivation of mentions combing all available context information.

The coding table was established with collaboration of four coders. 200 records were coded by three coders. The overall consistency rate is 85%. The other 200 records were coded by one coder.

Table 1. The list of policy document source institutions.

N.	Policy document source institution
1	UNESCO
2	The Publications Office of the European Union
3	International Union for Conservation of Nature
4	World Bank
5	World Meteorological Organization (WMO)
6	United Nations Environment Programme (UNEP)
7	Flemish Government Policy Documents
8	European Food Safety Authority
9	Intergovernmental Panel on Climate Change
10	Analysis & Policy Observatory (APO)

*Only 10 out of 40 institutions are presented here due to the space limitation

Results

Where are the mentions located?

Given the various types of policy documents, there is no fixed structure. However, based on the functional structure, locations of mentioning the scientific product can be summarized in Table 2.

Table 2. Location of mentioning scientific product in the policy document

Code	%	Code	%
1. Abstract	0.4%	7. Column	3.1%
2. Review	19.5%	8. Acknowledgment	0.4%
3. Methodology	3.4%	9. Appendix	3.1%
4. Expounding	53.4%	10. No mention	8.8%
5. Results	4.6%	11. Others	1.1%
6. Summary	2.3%	Total	100%

How are scientific products mentioned?

In total there are four forms in which scientific product could be mentioned in the policy document as shown in Table 3. Code *4. Non-standard reference* t only mention elements like title, DOI, or author of the scientific product.

Mentioned elements of scientific product are summarized in Table 4. Fragment entities (code 2.3) such as a number, or a chart are mentioned the most frequently. Sentences in conclusions (code 1.4) are also frequently used (23.7%).

Table 3. Different citing forms of scientific product in the policy document.

Code	%
1. Reference list	87.8%
2. Footnotes or endnotes	5.0%
3. Complementary material	5.0%
4. Non-standard reference	2.3%
Total	100%

Table 4. Mentioned elements of scientific product in the policy document.

Code	%
1. Sentences	30.9%
1.1. Title	0.8%
1.2. Abstract	1.1%
1.3. Method	4.2%
1.4. Conclusion	23.7%
1.5. Discussion	1.1%
2. Entities	35.5%
2.1. Opinion	5.7%
2.2. Concept	2.7%
2.3. Fragment	26.3%
2.4. Instrument	0.8%
3. Topics	21.4%
3.1. Theme	6.1%
3.2. Summary	7.3%
3.3. General	5.7%
3.4. Parallel reference	2.3%
4. Pure link	12.2%
Total	100%

Why are the scientific products mentioned?

Four major categories and twelve specific categories of motivations are identified, as shown in Table 5. It is highlighted that persuasive references (code 2) are dominant and have taken up 52% of total mentions. It means that scientific products are often used as evidence for justifying the argument proposed by the policy document.

Table 5. Motivation of mentioning scientific product in the policy document.

Code	%
1. Acknowledgeable reference	21.4%
1.1. Indicate the source of mentioned entities	10.7%
1.2. Indicate the source of mentioned examples	3.4%
1.3. Indicate the source of mentioned methodology	2.7%
1.4. Indicate the source of mentioned concept	2.3%
1.5. Indicate the source of mentioned topic	2.3%
2. Persuasive reference	51.5%
2.1. Support the argument by using examples	13.7%
2.2. Support the argument by listing relevant work	37.8%
3. Constructive reference	6.5%
3.1. Basis for reasoning	2.7%
3.2. Material of meta-analysis	3.8%
4. Informative reference	9.9%
4.1. Provide background information	6.1%
4.2. Provide complementary material	2.7%
4.3. Help to locate relevant studies	1.1%
5. Unable to judge	10.7%
5.1. Indirect reference	1.9%
5.2. Pure reference	8.8%
Total	100%

Conclusions

Main conclusions are drawn as follows. (1) Policy documents have seen most of the mentions of scientific product in the expounding parts (percentage = 53%) where core ideas are expressed, unlike scholarly articles where citations are mostly located in the literature review or introduction part. (2) Policy documents have mostly (percentage = 88%) followed academic citation style when mentioning scientific product, although it is not mandated. Meanwhile, there are other types of non-standard citing forms. (3) Among all types of mentioned elements of scientific product, fragment entities have been mentioned the most frequently, while conclusions of the scientific product are almost equally heavily mentioned. This indicates that policy documents are likely to mention the scientific product in a confirmative way, and could potentially be biased. (4) Over half of the mentions (percentage = 52%) are aimed to make the argument proposed by the policy document more persuasive. This indicates that policy document mentions of scientific product are indeed reflecting their societal impact because they are relevant in helping make decisions.

Compared with scholarly citations, policy document mention is likely to directly make use of the findings of the scientific product to inform practical guidance. To our observation, the policy document mention is separate from the academic discussion, therefore, no controversial, flattery, or critical mentions were involved. In this sense, policy document mention is more neutral.

Acknowledgments

The research is supported by National Natural Science Foundation of China (NO.71804067), Humanity and Social Science Foundation of Ministry of Education of China (18YJC870023), the Fundamental Research Funds for the Central Universities (No.30920021203) and Postgraduate Practice Innovation Project of Jiangsu Province (SJCX20_0135).

References

Bornmann, L., Haunschild, R. & Marx, W. (2016). Policy documents as sources for measuring societal impact. *Scientometrics*, 109(3), 1477-1495.

Kale, B., Siravuri, H. V., Alhoori, H. & Papka, M. (2017). Predicting research that will be cited in policy documents. *Proceedings of the 2017 ACM on Web Science Conference*, 389-390

Yu, H., Cao, X., Xiao, T. & Yang, Z. (2020). How accurate are policy document mentions? a first look at the role of altmetrics database. *Scientometrics*, 125(2), 1517-1540.

Is *Dimensions* a reliable data source of funding and funded publications?

Lin Zhang[1,2,3], Yinxin Zheng[1], Wenjing Zhao[1] and Ying Huang[1,2,3]

{linzhang1117, yancy_zheng, cady_zhao}@whu.edu.cn, ying.huang@kuleuven.be
[1] School of Information Management, Wuhan University, Wuhan 430072 (China)
[2] Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan 430072 (China)
[3] Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven B-3000 (Belgium)

Introduction

Launched by Digital Science in January 2018, *Dimensions* is a partly free scholarly database with the characteristic of linking and contextualizing different research information objects. As a young data provider, whether *Dimensions* is a reliable or even favorable data source has caused some attention in recent years. Thelwall et al. (2018) explored the coverage of Scopus journal articles in *Dimensions* and comparability between average citation counts in two databases; Herzog et al. (2020) assessed the coverage of publications from *Dimensions* in comparison with other data sources. Less attention so far has been given to examining the coverage and accuracy of funding and the funded publications in *Dimensions*, and this gap in literature leads to our study. In this poster, we present the following preliminary results:

1) the completeness of funding data in *Dimensions*, compared to the related funder's official website;

2) a comparison between *Dimensions'* collection of funded publication data and those provided by the Web of Science (WoS).

Data

National Natural Science Foundation of China (NSFC), as the largest funding agency for basic research in China, is selected as the primary object for comparison due to its unique position in the Chinese scientific funding system. The number of funded projects provided by the statistics report on the NSFC official website[1] is used as a baseline for assessing the coverage of *Dimensions'* grants data. Further, the seven major types of funded projects from 8 scientific departments (2001-2018) in the NSFC are selected for comparison in the current study. A unified 8-digit NSFC grant number that contains unique information to distinguish project type, application year and scientific department is used to match the project data. In total, 358,287 and 404,976 funded projects applied during 2001-2018 were obtained from *Dimensions* on Oct. 2019 and Jan. 2021, respectively. As for the baseline, 407,076 funded projects from NSFC are used for comparison.

In terms of funded publications, the grant number of funded projects completed during 2001-2018 and publicized in the NSFC official platform[2] is used to retrieve the related publication data from both *Dimensions* and WoS. The total amount of funded publications obtained are 2,068,242 (*Dimensions*) and 2,414,222 (WoS) respectively.

Results

Funded projects

As seen in Figure 1, the number of NSFC-funded projects retrieved from *Dimensions* respectively in Oct. 2019 and Jan. 2021 are in general identical, with the only exception of the year 2006. The funded project data for 2006 obtained in 2019 has a striking deficiency in *Dimensions*, compared to that from NSFC's website. However, the data for 2006 was improved largely in *Dimensions* by Jan. 2021, in spite of a small degree of incompleteness. Although part of the data for 2018 has not been updated timely in *Dimensions*, it seems the coverage of data in *Dimensions* is continuously improving.

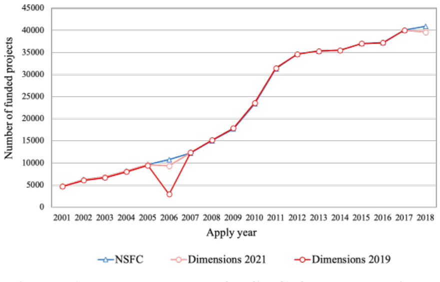


Figure 1. The number of NSFC-funded projects

To provide a more refined analysis, we calculate the coverage ratio of projects in *Dimensions* for seven different funding types and eight scientific departments in NSFC. In particular, the number of projects in *Dimensions* (Jan. 2021) is divided by the number of projects obtained from the NSFC report. As illustrated by Figure 2, the number of projects of GP, YS, EYS, LDR and DYS types are in general consistent with the number obtained from the NSFC

[1] www.nsfc.gov.cn/publish/portal0/tab505/

[2] http://output.nsfc.gov.cn

except for 2006. Notably, the coverage of KP projects in the *Dimensions* is only 56% in 2006. Further, different degrees of deficiency for JP projects can also be observed in most years. The varying degrees of missing data for all types of projects in 2018 might be due to Dimensions' update issue.

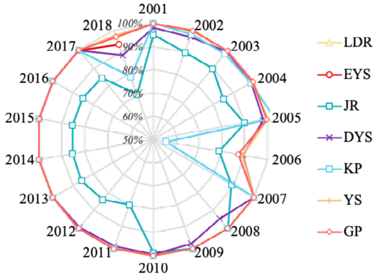


Figure 2. Coverage of 7 types of funded projects

Note: LDR= Fund for Less Developed Regions, EYS= Excellent Young Scientists Fund, JR= Joint Research Fund for Overseas Chinese Scholars and Scholars in Hong Kong and Macao, DYS= National Science Fund for Distinguished Young Scholars, KP= Key Program, YS= Young Scientists Fund, GP= General Program

Similarly, Figure 3 indicates the exceptional low coverage for projects in all departments in 2006 and 2018. In terms of different departments, the numbers of projects in CS, LS, ES, MS and HS are basically identical with that from the NSFC report, while the numbers in M&PS, E&MS and IS deviate from the official records in NSFC.

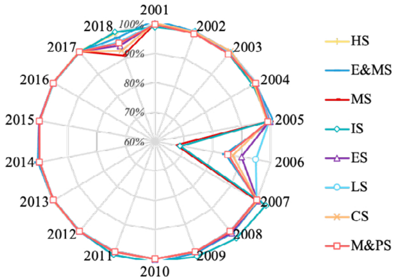


Figure 3. Coverage of funded projects in 8 scientific departments

Note: HS= Health Sciences, E&MS= Engineering and Materials Sciences, MS= Management Sciences, IS= Information Sciences, ES= Earth Sciences, LS= Life Sciences, CS= Chemical Sciences, M&PS= Mathematical and Physical Sciences

Funded publications

From Figure 4, a significant increasing trend can be observed in both the total and the average number of publications indexed in two databases. For the average number of publications, *Dimensions* took a leading position before 2008 but was overtaken by WoS since 2009. In addition, the remarkable contrast of the average number of publications in 2009 between two databases might indicate a certain level of publication data deficiency in *Dimensions*. The relatively low number of publications in 2014 in both databases is mainly attributed to the low number of funded projects in that year.

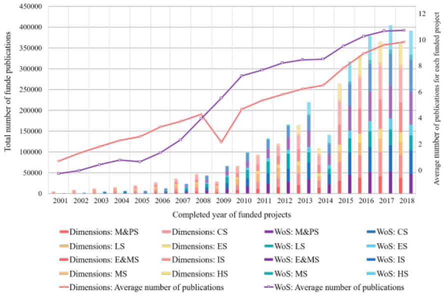


Figure 4. The total and average number of publications for funded projects

Discussion and Conclusions

In general, *Dimensions* provides accessible and relatively reliable data on funding (at least for NSFC-funded projects). However, the data coverage may vary with different years and project types, and a delay of data updates is observed for recent years. For the funded publications, data retrieved from *Dimensions* and WoS shows clear divergence. Since the completeness and accuracy of funding information derived from publications in bibliographic databases are often questioned by scholars (Álvarez-Bornstein et al., 2017), whether *Dimensions* provides satisfactory or complementary information remains to be further explored.

In subsequent research, both the funded projects and publications in different databases will be matched precisely based on the grant number of projects and DOI, respectively, for a refined analysis on the data accuracy of *Dimensions*. Furthermore, other official sources of funded projects and publications, such as the US National Science Foundation (NSF), will be included to examine the data integrity of *Dimensions* from a more comprehensive perspective.

References

Álvarez-Bornstein, B., Morillo, F. & Bordons, M. (2017). Funding acknowledgments in the Web of Science: completeness and accuracy of collected data. *Scientometrics,* 112(3), 1793-1812.

Herzog, C., Hook, D. & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies,* 1(1), 387-395.

Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science?. *Journal of informetrics,* 12(2), 430-435.

The effect of international academic mobility on international cooperation: from the perspective of overseas institutions

Zhang Ying[1,2], Liu Xiaomin[1,2] and Sun Yuan[3]

{zhangying, liuxm}@mail.las.ac.cn, yuan@nii.ac.jp
[1] National Science Library, Chinese Academy of Sciences, Beijing (China)
[2] Department of Library, Information and Archives Management, School of Economics and Management, UCAS, Beijing (China)
[3] National Institute of Informatics, Tokyo (Japan)

Introduction

This study explores how the mobility of researchers among international institutions affects their international collaboration. We took 254 researchers who returned to China from overseas as the samples, then established panel data through personal homepages and Scopus database. Finally, we utilized two-way fixed effect model (FE model) and random effect model (RE model) to analyze the relationship between mobility and international collaboration.

Data & Methods

In the study, we took researchers who returned to Chinese Academy of Sciences between 2011 and 2015 as samples. After excluding those without complete information, there were 254 researchers included. 48.4% of them earned the PhD overseas, then worked abroad for at least 3 years as postdoctor or formal staff. The rest got PhD in China, and they worked at least 5 years overseas after that.

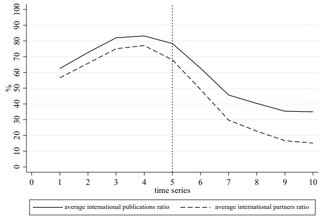
We established panel data with a time series of 10 years, including data 5 years before and 5 years after their return. Personal data (age, gender, discipline, education and work experience) were from personal homepages; publication data were collected through Scopus based on the author unique identifier. The variables and their descriptions are shown in Table 1. A publication was defined as an international publication if its authors' institutions belong to different countries. An author was regarded as an international partner if whose institution do not belong to the same country with our target researchers (the 254 samples).

As for the analysis model, the hausman-test showed that the FE model was more appropriate than RE model. In order to simultaneously control the individual effect of researchers and time effect, we constructed two-way FE model. But in order to reveal the effects of the time-invariant variables, we still further ran a RE model. Finally, we conducted the test of robustness by dividing the samples into two parts according to whether they have overseas degrees, which proved that our results were stable and reliable. (The robustness test results are not shown because of poster's space limit).

Table 1. Descriptions of variables

Variables	Descriptions
Dependent variables	
international publications ratio$_{it}$	ratio of international publications to total publications published by researcher i in year t
international partners ratio$_{it}$	ratio of international partners to total partners of researcher i in year t
Explanatory variables	
length of time spent overseas$_{it}$	number of years spent in overseas institutions by researcher i until year t
number of overseas institutions$_{it}$	number of institutions experienced by researcher i until year t
postdoctor$_i$	1 if researcher i had overseas postdoctoral experience; 0 otherwise
formal position$_i$	1 if researcher i had overseas formal working position; 0 otherwise
Control variables	
age$_{it}$	age of researcher i in year t
gender$_i$	1 if male; 0 otherwise
discipline_1$_i$	1 if the discipline is fundamental frontier interdisciplinary sciences; 0 otherwise
discipline_2$_i$	1 if the discipline is resources ecology and environment sciences; 0 otherwise
discipline_3$_i$	1 if the discipline is life and health sciences; 0 otherwise
discipline_4$_i$	1 if the discipline is advanced materials sciences; 0 otherwise
discipline_5$_i$	1 if the discipline is optoelectronics and space sciences; 0 otherwise
America$_i$	1 if researcher i visited America; 0 otherwise
Germany$_i$	1 if researcher i visited Germany; 0 otherwise
Japan$_i$	1 if researcher i visited Japan; 0 otherwise
Britain$_i$	1 if researcher i visited Britain; 0 otherwise
other countries$_i$	1 if researcher i visited other countries; 0 otherwise

Results & Discussion



When "time series"=5, the researchers returned to China

Figure 1. Average international cooperation in the 5 years before and 5 years after returning

From the results of descriptive analysis, researchers' average ratios of international publications and partners had shown a similar trend. In the period of 5 years before the return, both of them had increased steadily from about 60% to about 80%. But during the 5 years after their return, they had dropped sharply to about 35% and 15% respectively.

Table 2. Results of two-way FE panel regression

Variables	international publications ratio	international partners ratio
length of time spent overseas$_{it}$	0.131***	0.140***
	(0.015)	(0.013)
number of overseas institutions$_{it}$	-0.103***	-0.102***
	(0.026)	(0.023)
age$_{it}$	-0.083***	-0.101***
	(0.009)	(0.008)
_cons	2.608***	3.113***
	(0.298)	(0.268)
N	2540	2540
r2	0.214	0.351

Standard errors in parentheses=
"*p < 0.05, ** p < 0.01, *** p < 0.001"
Note: All variables were taken into account in the model, but here we only presented the results of time-varying variables. And because our data included the 5 years before researchers' return, the length of overseas time varied over time before the return, so did the number of overseas institutions.

When other variables were controlled, length of overseas time and number of overseas institutions had opposite effects on international cooperation. An extra year in overseas boosted about 14% in both ratios of international publications and partners. However, the two ratios both decreased by about 10% for each additional overseas institution experience. Meanwhile, as the age increased by a year, international publications and partners ratios decreased by 8.3% and 10.1% respectively.

The results of RE model have confirmed our previous results of time length, institution number and age. Moreover, when other variables were controlled, postdoctoral experience could boost the two ratios by about 9%. But formal position did not show significant effect on them. As the research fields, only discipline_5 showed significantly better than discipline_1 in international cooperation. However, both the gender and countries visited had no significant effects. As for the negative impact of the number of overseas institutions, an additional robustness test was carried out. We controlled for the researchers' productivity and international network, then the results cleared that the negative impact of the number of overseas institutions on international cooperation still remained significant.

Table 3. Results of RE panel regression

Variables	international publications ratio	international partners ratio
length of time spent overseas$_{it}$	0.023***	0.029***
	(0.005)	(0.005)
number of overseas institutions$_{it}$	-0.055***	-0.066***
	(0.015)	(0.014)
postdoctor$_{i}$	0.082*	0.090**
	(0.036)	(0.033)
formal position$_{i}$	0.043	0.043
	(0.025)	(0.023)
age$_{it}$	-0.030***	-0.030***
	(0.005)	(0.005)
discipline_5$_{i}$	0.113**	0.075*
	(0.039)	(0.036)
_cons	1.380***	1.321***
	(0.161)	(0.148)
N	2540	2540

Standard errors in parentheses=
"*p < 0.05, ** p < 0.01, *** p < 0.001"
Note: To avoid collinearity of disciplines, we chose discipline_1 as reference variable. Due to space limit, Table 3 only showed the results of explanatory variables and some control variables which had significant influence.

Conclusion

To build better international cooperative relations, increasing the length of time spent abroad is desirable; however, frequent movement between different institutions is discouraged. As for the type of work, post-doctoral work is beneficial to international cooperation, but the effect of formal work is not significant. In addition, people tend to cooperate with people who are younger, regardless of their gender. And it makes no difference which country you choose to stay in.

This study provides evidence from China on the impact of international mobility on cooperation and can provide reference for researchers to make overseas work plan. Next, we will further improve this study by heckman two-stage method to solve the endogeneity problem caused by sample bias.

Research on international patent technology layout in high-speed railway field

Rongying Zhao[1], Ruru Chang[2], Xiaoyu Wang[3] and Zhaoyang Zhang[4]

[1] *zhaorongying@126.com*
[2] *Changrrrr@whu.edu.cn*
[3] *1942679056@qq.com*
[4] *windboy727@vip.qq.com*
School of Information Management, Wuhan University, Wuhan, Hubei Province (P.R.China)

Introduction

The development and mastery of railway technology provide a solid foundation for economic and social development. The patent layout is for targeted improvement of patent layout strategy. Therefore, itis particularly important to fully understand technological development and market competitive advantages from the overall impact of high-speed rail on the regional city economy. We observe the trend of patent distribution from three perspectives: vertical, horizontal and depth. Based on the comprehensive analysis of the patent technology layout in the field of high-speed rail, this paper further analyses the patent layout of the main competitors. A scientific patent measurement is proposed to accurately identify market gaps and major competitors. At the same time, by examining the development of technology competitiveness through patenting as undertaken to provide useful insights.

Data and Method

The 12,060 patents in this paper are obtained from the Derwent Innovations Index (DII) by subject search. One is technology classification, subject related to high-speed rail technology can be roughly divided into five categories as shown in Table 1. The other is special appellation in various countries, such as Train a Grande Vitesse (French), Alta Velocidad Española and talgo (Spanish), Korea Train Express (Korea), Acela Express (American), Inter City Expressand and Transrapid and Eurostar (German and European), ETR (Italian), High-Speed Line (Belgium), LRC (Canada), Shinkansen (Japan), Pendolino (British), ICN (Swiss), Sokol (Russian), Intercity Express and CRH and THSR (China) (Dang, 2016; Fumihiro Hasegawa et al., 2018).

Patent layout refers to the overall layout behaviour of the company's own patent situation. Patent layout, not only must always pay attention to the quantity, but also monopolize the technology market through patent portfolios, prevent competitors from attacking their own technical weaknesses, and avoid falling into infringement disputes. The layout content includes the total number of patent applications, the fields involved in the patent application, the region to which the patent application is applied, and the useful life of the patent application. It can be summarized as follows: the time, place, type and quantity of patents in general.

Table 1. The categories of high-speed rail technology patent terms

Categories	Terms	Results
High-speed train/train	bullet train; high-speed train; high-speed rail*; express railway	9,082
EMU	train-set; motor train set; rail motor car*; multiple-unit train	377
High-speed railway	rapid transit railway	244
Express train	Express train	93
Maglev train	maglev train; Magnetically Levitated Train; magnetic suspension train; JR-Maglev	982

Results and Discussion

Vertical analysis: the changes of patent applications

The global high-speed rail technology patent application trend can show the development history of high-speed rail technology to a certain extent. The patent application date is closer to the invention time of the patent. In this paper, the application date is used as the time basis for data statistics.

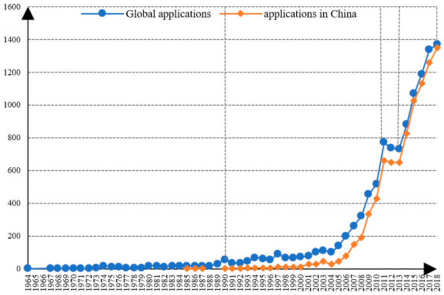


Figure 1. The annual change trend of patent applications for high-speed rail

It can be roughly divided into four stages: the initial period (1964-1990), the rising period (1991-2010), and the adjustment period (2011-2013), the rapid development period (2014 to present).

With the rapid development of computer technology, high-speed rail technology will face more changes and progress. At the same time, according to the development of high-speed rail technology, the closer global trade cooperation will promote the optimization and innovation of high-speed rail technology.

Horizontal analysis: patent sources and geographical distribution

By analysing the country of the patent applicant, the main technology source country in this technical field can be identified, and the patent layout of the high-speed rail patent market can be clarified, See Figure 2.

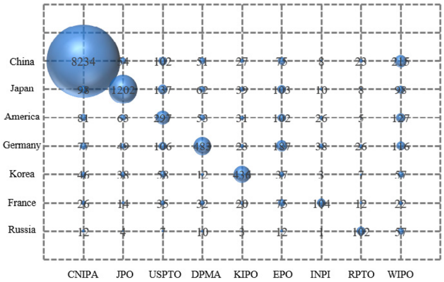


Figure 2. Patent sources and geographical distribution in the high-speed rail field

Most of countries are focusing more on the patent layout in their own countries. Especially for China, its domestic layout is much larger than its international layout. In overseas markets, China pays more attention to the United States and Europe.

In addition, other countries also pay more attention to the Chinese market and actively carry out patent layout. China is the third largest target country for Japan, the US and Germany besides themselves, and the second largest target country for South Korea and Russia's overseas deployment. The international distribution is much higher than that of most countries in Japan. It can be seen that Japan attaches great importance to the overseas patent technology market.

Depth analysis: major global patent technology companies

In order to more specifically understand the future development direction, it is necessary to further analyse the main patentees. Patent quantity and patent quality are two factors to measure competitive advantage. Patent quantity refers to patents held by patentees (enterprises) and patent quality is an index related to patent citation. We use average patent citation rate as a measure of patent quality. Taking the patent applications as the abscissa and the patents citation as the ordinate, and the larger the bubble, the larger the total number of patents cited, See Fig. 3.

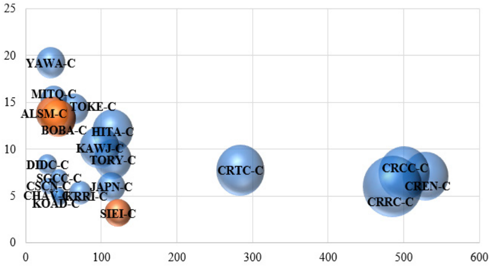


Figure 3. Bubble chart of top 20 enterprises in high-speed rail field

The Nippon Steel & Sumitomo Metal Corporation has highest average patents citation rate (19.21%). Most of these 20 enterprises are Asian enterprises, including seven Chinese enterprises, one Sino foreign joint venture and seven Japanese enterprises. As a big high-speed rail country, China has shown strong momentum in patent applications in the field of high-speed rail, and the number of patent applications is far ahead. However, in terms of patent quality, Chinese enterprises still need to improve.

Conclusions

This paper makes a macro analysis of the patented technology in the high-speed rail field from three dimensions: vertical, horizontal and depth. It is expected that patent applications in the high-speed rail field will continue to maintain a rapid growth trend in the future. Then, the awareness of international protection of patented technology needs to be improved from country to country, especially in China. Finally, patent quality can't be ignored, while enterprises keep the advantage of high patent output. Both countries and enterprises should reduce redundant patent applications and strengthen technological innovation to improve patent quality.

Acknowledgments

This research is funded by the National Social Science Found Major Project of China (18ZDA325).

References

Dang, X. J. (2016). *Patent Early Warning of Chinese High-speed Railway Technology* (Master's thesis, Beijing University of Technology).https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201602&filename=1016717513.nh

Hasegawa, F., Taira, A., … & Kim, I. (2018). High-speed train communications standardization in 3gpp 5g nr. *IEEE Communications Standards Magazine*, 2(1), 44-52.

A Bipartite Author-algorithm Network for Exploring Versatile Scholars

Yi Zhao[1], Yuzhuo Wang[2] and Chengzhi Zhang[3]

{[1]yizhao93, [2]wangyz, [3]zhangcz}@njust.edu.cn
Nanjing University of Science and Technology, 200 Xiaolingwei Street, Nanjing 210094 (China)

Introduction

The need for particular skills is one of the factors that motivate scholars to find collaborators. However, it's difficult to know directly what fine-grained skills (e.g., algorithm) a scholar masters, but this kind of information is deeply embedded in scholars' publications, especially in the artificial intelligence domain. For example, if a researcher wants to use alternating directions dual decomposition (AD^3) algorithm to address a natural language processing (NLP) problem, he/she may want to know the following information: Which researchers are studying the AD^3 algorithm, and with whom could the researcher collaborate? Even researchers who study AD^3 prefer to publish solo-authored publications; code sharing is ubiquitous in the computer science field, and we can also reproduce their algorithm via the codes they provided in their paper. Therefore, it is essential to build a connection between the authors and the algorithms, and the network can help researchers find suitable collaborators who master particular algorithms. In this paper, we briefly 1) construct a bipartite author-algorithm network and 2) explore the scholars with the most diverse skills (i.e., *versatile scholars*) in the NLP domain, versatile scholars are those who mentioned a variety of algorithms in their papers.

Methodology

Dataset

Our dataset consists of two parts: algorithm entities and author information. For the algorithm entities, we used the in-house dataset annotated by Wang and Zhang (2020). To obtain author information, we first downloaded and parsed the paper title, author name, paper ID, publication year from the dataset (https://github.com/lingo-iitgn/NLPExplorer), which was built using the data provided in the ACL Anthology (https://www.aclweb.org/anthology/). Then, we extracted authors' affiliations, affiliation locations from each paper manually. Finally, we integrated the two parts through paper IDs provided by the ACL Anthology. Moreover, the ACL Anthology volunteer team has used multiple approaches to address name ambiguities (Mohammad, 2020).

Method

Bipartite network is a graph on a node set which represents two different types of entities and relational ties occur only between nodes of different types. A bipartite network is represented as follows:

$$\mathrm{G} = (\mathrm{U}, \mathrm{V}, \mathrm{E}) \qquad (1)$$

Where U and V are two types of nodes in network G. U and V denote the authors and algorithms respectively. E is the set of edges in G.

The indicator of *effective partners* (EPs) was used to explore most versatile scholars (Bersier et al. 2002). In formulae:

$$H_j = -\sum_{i=1}^{s} \frac{a_{ij}}{a_{\cdot j}} log_2 \frac{a_{ij}}{a_{\cdot j}} \qquad (2)$$

$$EPs_j = \begin{cases} 2^{H_j} \\ 0, if\ a_{\cdot j} = 0 \end{cases} \qquad (3)$$

Where a_{ij} is the number of co-occurrences between author j and algorithm i, $a_{\cdot j}$ is the total number of co-occurrences between author j and all algorithms. Eps_j is the effective number of algorithms mentioned by author j.

Results

Statistics of Dataset

Table 1. presents the basic statistical description of our dataset. 891 types of algorithm entities are extracted from ACL papers, and 5,619 authors published ACL papers between 1979 and 2015.

Table 1. Statistics of the dataset

	author	affiliation	country	algorithm
total	12,142	12,711	12,711	59,277
unique	5,619	1,034	65	891

The bipartite network of authors and algorithm entities in the NLP domain

We choose eight as the threshold of the minimum number of author-algorithm co-occurrences to generate better visualization. Fig.1 displays the bipartite network. The author node is red, and the algorithm node is blue. A link between the author and the algorithm exists if the author has published a paper that mentioned the algorithm. The thickness of the link represents the number of papers an author published that mentioned an algorithm. The size of the node denotes the number of papers an author published (red) or the number of papers published that mentioned an algorithm (blue).

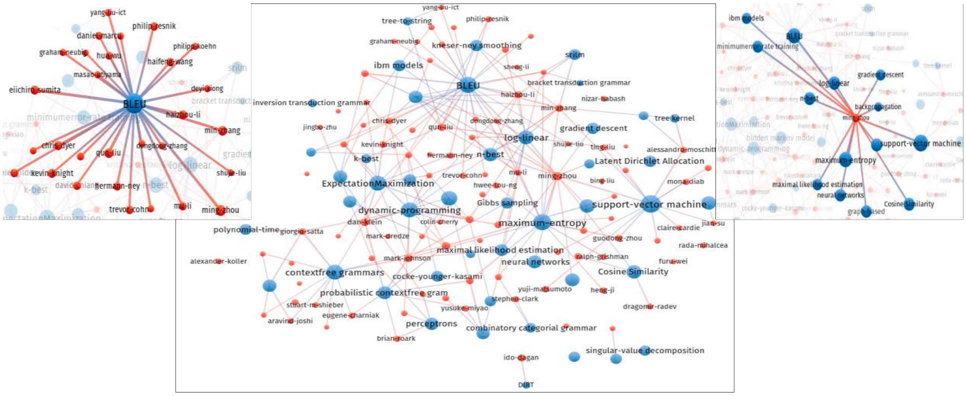


Figure 1. Bipartite network of authors and algorithm entities (tool: VOSviewer)

As Fig.1 shows, support vector machine, BLEU and maximum entropy are widely mentioned by authors. In addition, almost all the blue nodes are much larger than the red nodes. 4.52 types of algorithms are mentioned per paper, whereas the average number of publications per author is 2.14. BLEU is widely used for the evaluation of the results of machine translation, it was mentioned by 29 authors (Fig.1, upper left). For Ming Zhou, a famous NLP scholar, 117 types of algorithms are extracted from his papers (Only those with frequency over 8 are shown in Fig.1, upper right). According to our statistics, BLEU (proposed by Papineni, etc.) and maximum-entropy (proposed by Jaynes) are his two most mentioned algorithms, spanning 17 and 15 papers, respectively, and it may suggest that he is adept at using these two algorithms. Additionally, machine translation is perhaps one of his research interests.

The most versatile scholars in the NLP domain

We calculated the EPs and explored who is the most versatile scholar in the NLP domain, as shown in Table 2.

Table 2. The top-5 versatile scholars

Author	EPs	Affiliation	Location
Christopher Manning	83.34	Stanford University	USA
Ming Zhou	77.76	Microsoft Research Asia	China
Noah Smith	76.96	University of Washington	USA
Daniel Klein	75.54	University of California at Berkeley	USA
Chris Dyer	65.03	Carnegie Mellon University	USA

Eps provides interesting rankings for authors who have mentioned or even used a variety of algorithms in their papers. Most of the top-5 versatile scholars are influential scholars in the NLP domain: ACL Fellows (i.e., Christopher Manning, Noah Smith), ACL president (i.e., Ming Zhou). Four of the five authors are affiliated with elite universities in the USA, while Ming Zhou is affiliated with an excellent firm that is located in China. It may indicate that in terms of algorithm usage, American scholars still dominate the NLP community.

Conclusion

This paper is intended to build a bipartite author-algorithm network to facilitate the development of applications, such as finding collaborators, exploring scholar's algorithm usage. The author-algorithm network displays that each author has mentioned what types of algorithms in their papers. Moreover, we found that 4 of the 5 versatile scholars are from the US. Given that whether the algorithm is truly used by the author has not been considered, more types of models should be used to classify the algorithm citation functions. The validity of the author-algorithm network to facilitate the scientific collaboration will be demonstrated in the future.

Acknowledgments

This work is supported by National Social Science Fund (No. 18BTQ073), Jiangsu Social Science Fund (No. 20TQA001), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX20_0347).

References

Bersier, L. F., Banassek-Richter, C. & Cattin, M. F. (2002). Quantitative descriptors of food-web matrices. *Ecology,* 83(9), 2394-2407.

Mohammad, S. M. (2020). Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp.7860-7870). Online: ACL.

Wang, Y. & Zhang, C. (2020). Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. *Journal of Informetrics,* 14(4), 101091.

Citation Time Entity Recognition Based on Deep Learning

Yufei Zhu[1], Si Shen[2] and Dongbo Wang[3]

{[1] zhuyufei, [2] shensi}@njust.edu.cn
Nanjing University of Science and Technology (China)

[3] db.wang@njau.edu.cn
Nanjing Agricultural University (China)

Introduction

With the increasing enrichment of academic resources, some scholars have been beginning to study how to improve the utilization rate of academic resources, such as mining the content of academic full text. Academic full text contains a lot of important academic resource information. Extracting the key points in academic full text by means of entity annotation and recognition can facilitate further analysis and research of text content. For example, building an academic resources network by the creation of knowledge mapping technology. These work make the visualization of academic information, the integration of academic resources, easy to retrieve and query, and improve the efficiency of academic resources.

Some scholars have proposed the concepts of publication year, date entity when constructing the ontology of scientific papers, but most of them explore the time association between papers from the perspective of papers, rather than focusing on the content of academic full text. This paper analyses the citation content in the academic full text at the median level, and the concept of citation time entity is proposed. Citation time is the time when the citation is published in the academic full text. The citation time has been listed in the references of the journal, but only the author, institution and literature name are listed together with it. If the citation time is annotated and extracted in the full text, the cited content can be annotated and extracted at the same time. The cited contents include the research problems, research methods and results of citations. When the citation time is marked and associated with the quoted content, we can build a dynamic topic model and other technologies by considering the time factor to analyse the research trends, research topics, citation authors' writing trends, and the changes of tools and models used in the research in recent years, so as to further tap the resources in the academic full text. The logical framework of association between citation time and cited content in academic full text is shown in Figure 1.

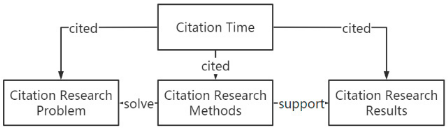

Citation Time
cited
cited
cited
Citation Research Problem
solve
Citation Research Methods
support
Citation Research Results

Figure 1. The citation time is associated to the quoted content.

In this paper, scientific literature published in JASIST journals in recent three years is selected for the citation entity tagging experiment by using the method of sequence tagging of CRF and BERT models. The experiment of named entity recognition of citation time is carried out from academic full text.

Data Sets and Methods

This paper selects 502 papers published in JASIST journals from 2017 to 2020, and finds examples of citation time in the literature, and then unifies the labelling specification. The entity introduction and examples are shown in Table 1.

Table 1. Introduction of citation time entity.

Name	Introduction	Example 1	Example 2
Citation time	The publication time of citations in academic full text.	In **2003** {ref[#asi 22654-bib-0024]} , Jansen and Spin.	By **1994** the confusion within the field had already.

After that, we selected 15 taggers to annotate the citation time entities on the BRAT platform, and the results are shown in Table 2. In this experiment, a total of 3876 citations were annotated, of which 3873 were 1 in length, that is, most of the citations were in a single year, such as 2016, 2017, etc.

Table 2. Statistics of entity annotation.

No.	Introduction	Count
1	Paper	502
2	Citation time entity	3876
3	Average citation time entity	7.72
4	Citation time entity of length 1	3873
5	Citation time entities with length greater than 1	3

We find that there are few types of citation time entities, that is, most of them are specific years in the past 20 years. The frequency statistics of each year is shown in Figure 2.

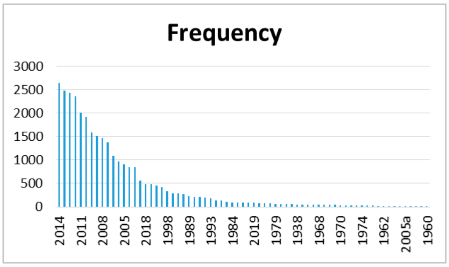


Figure 2. Statistics of citation time frequency distribution.

Conditional random fields (CRF) was first proposed for sequence data analysis, and has been successfully applied to named entity recognition in natural language processing. In this paper, CRF + + is used to recognize the named entity of citation time.

BERT is a very powerful model in NLP field, and can be used to solve the problem of sequence annotation. In this paper, we use BERT to do named entity recognition experiments on citation time entities.

SCIBERT is a BERT model trained in scientific writing. It has its own vocabulary (SCIVACAB), which is most suitable for training corpus.

Results

When we using the CRF + +, the hyper parameter is set to 10.0, and the threshold of training feature is set to 10. In the training of BERT model, a training model with 12 layers, 768 hidden units and 12 self attention heads are selected. The crossing data set is set up to be trained for 6 times by the BERT model. Finally, the data set is trained by SCIBERT once.

The evaluation effect of the model is shown in Table 3.

Table 3. Evaluation of model effect.

Model	Accuracy	Precision	Recall	F1
CRF	0.9981	0.8652	0.9555	0.9082
BERT-1	0.9996	0.9861	0.9775	0.9818
BERT-2	0.9997	0.9889	0.9816	0.9853
BERT-3	0.9997	0.9909	0.9788	0.9848
BERT-4	0.9998	0.9883	0.9922	0.9902
BERT-5	0.9997	0.9925	0.9771	0.9847
BERT-6	0.9996	0.9768	0.9880	0.9824
SCIBERT	0.9996	0.9739	0.9713	0.9726
MIN	0.9981	0.8652	0.9555	0.9082
MAX	0.9998	0.9909	0.9922	0.9902

From the experimental results, it can be seen that the effect of the three models for named entity recognition is good, and the evaluation effect of BERT model is the best, which can achieve the accuracy of 0.9998. The evaluation effect of CRF is the worst, its evaluation values are all the lowest among the four indexes in this experiments. The evaluation effect of SCIBERT is between two other models, but there is little different from BERT. The citation time entities in the content of academic papers are easy to train because they are relatively standardized, most of them are the specific years with length of 1, and there are few categories of citation time entities. It can be seen that citation time is suitable for automatic and semi-automatic entity annotation on large-scale corpus.

Conclusion

This paper puts forward the concept of citation time entity, which is often associated with citation problems, citation research methods, citation results and other content. We can use time topic model and other technologies to analyse the changes of academic content in recent years, such as research trends, research topics, tools and models used in research, so as to fully tap the resources in academic full text. This paper also marks the citation time of 502 scientific papers published in JASIST in recent four years. We found that the citation time format of academic papers is very standardized, most of them are individual years. Then we use the CRF and BERT models to do named entity recognition experiments on annotated corpus, and evaluate the experimental results. The experimental results show that the effect of citation time on entity recognition is better, and it is concluded that citation time is very suitable for automatic and semi-automatic entity annotation in large-scale corpus. In the future research, we can do further research on the reference resolution of citation time, or on the topic evolution of time factors by integrating citation research problems, citation research methods and citation authors.

Acknowledgments

The authors acknowledge the National Natural Science Foundation of China (Grant Numbers: 71974094) for financial support.

References

Shu, L., Xu, H. & Liu, B. (2017). Lifelong Learning CRF for Supervised Aspect Extraction. ACL.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.

AUTHOR INDEX

SPONSORS

DIAMOND

International Center for the Study of Research

GOLD

DIGITAL science

RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

SILVER

JOURNAL OF DATA AND INFORMATION SCIENCE

JDIS 数据与情报科学学报

Clarivate™

QSS Quantitative Science Studies

BRONZE

frontiers

SPRINGER NATURE

WITH THE SUPPORT OF

fwo

ISSN 2175-1935

9 789080 328228