Methodological Procedure to Overcome the Lack of Normalisation of Author Names in Bibliometric Analyses at the Micro Level

Rodrigo Costas and María Bordons

rodrigo.costas@cindoc.csic.es, mbordons@cindoc.csic.es,
Centro de Información y Documentación Científica (CINDOC), CSIC, C/Joaquin Costa 22,
28002 Madrid (Spain)

Introduction

Obtaining scientific production of individual scientists from bibliographic databases raises the challenge of how to deal with the lack of normalisation of author names. The fact that names of scientists are not normalised in bibliographic databases has been pointed out by different authors as a serious limitation in the bibliometric analyses at the micro level, since it significantly reduces the accuracy of production-based measures (Wooding et al., 2004; Ruiz-Pérez et al., 2002). In this paper, a new methodology to solve the problem of lack of normalisation in the author field of bibliographic databases in the development of bibliometric studies is presented.

Objectives

Our aim is to develop a method to obtain the scientific production of individual scientists overcoming difficulties due to signing variations and homonyms. We focus on problems derived from the use of Thomson-ISI databases, although our procedure could be adapted to other bibliographic databases. The methodology is applied here to the study of the scientific production of permanent scientists in the Natural Resources area at the Spanish Research Council (CSIC) during the years 1994-2004.

Methodology

Sources

4. Full name and working centre of 333 researchers with a permanent position at CSIC, Natural Resources area, were provided by the institution in a personnel file.
5. Scientific publications were obtained from the Science Citation Index CD-ROM, years 1994-2001. Two different methods were used in the search strategy: a) search by address: documents signed by any Natural Resources' centre (9,109 documents); and b) search by author: documents signed by any of the CSIC scientists if they had entered CSIC between 1994 and 2001 (1,807 documents). The latter search was used to include in the study the production of recently-incorporated scientists.

Author identification

The procedure here presented comprises four different steps:

6. Obtaining raw data of productivity of scientists. To avoid confusion due to homonyms, an "Author-Centre" table was created, in which every signing author is assigned to one working centre (following Bordons et al., 1995) and productivity for each entry "author-centre" was calculated.
7. Building a list of name variants. The "basic structure" of the original names of our researchers (following Ruiz Perez et al, 2002) is considered as follows: First-Surname Second-Surname, First-Name [Middle-Name]. For example: "Garcia Casas, Jose [Luis]". Nine different name variants were created for every scientist, including all the possible ways of signing documents for each scientist.
8. Identification of signing variants in the file of scientific output. The list of potential name variants was matched with the real signing forms included in the "Author-Centre" table and every match was marked with a personal identification code. Two different types of signing forms were found: name variants (more than one name for a scientist) and institutional variants (more than one institution assigned to a specific scientist).
9. Final cleansing. Obtaining final productivity of scientists. Two different types of author-centre entries marked with a personal identification code were obtained: a) entries automatically assigned to studied researchers, since both author name and centre matched the data found in the personnel file; b) entries whose assignation needed to be confirmed, because either the name or the centre differed from that of the personnel file. Scientific production of every automatically identified author was compared to that signed by its potential variant signing form: a high degree of coincidence in co-authors, affiliation institutions, publications journals and article titles supported the hypothesis of being the same scientist. Remaining doubts about the identity of specific

scientists were resolved through Internet searches and experts' advice.

Results

The name of 299 scientists (90% of the scientists) followed the basic structure and their production was obtained according to the procedure here shown. Authors with no basic structure were handled manually. Around 92% of all the scientists had at least one ISI document in the period.

Around 18% of original authors had more than one variant name, which ranged from 2 to 3 different names per author. In relation to institutional mobility, 67% of the authors had a single institution in the period, while 28% showed 2 or more different institutions.

Considering both name and institutional variants, 60% of authors showed just one signing form, while 30% showed 2 signing forms either due to name or institutional variants.

A total of 19,349 author-centre entries were automatically included in the original "Author-Centre" table. After matching personnel data to the list of potential name variants, 1,176 entries were assigned to the studied scientists. It means that 94% of original entries were automatically discarded, only entries with any chance of belonging to any of the studied researchers being analysed.

Table 1. Global results of the identification of authors.

Assigned signing forms	**Total**	**%**
Automatically	257	21.85
To be revised	919	78.15
Total	1176	100
Revised signing forms		
Positively revised	217	23.61
Discarded	702	76.39
Total	919	100
Total signing forms assigned	474	

As we can see in Table 1, a total of 257 author-centre entries were automatically and definitely assigned to the 333 studied scientists (22% of total signing forms assigned), while 919 entries (78%) needed to be revised, of which 24% were then positively assessed and accepted. In the end, a total of 474 author-centre entries were definitely assigned to the studied scientists (40% of initially assigned signing forms). The scientific publications finally assigned to the studied scientists amounted to 3,302 documents.

Conclusions

Our methodology requires the "basic structure" of author names as a starting point, but this structure is the norm, as shown by the fact that 90% of the CSIC scientists studied here followed it.

Although 82% of the scientists always signed with the same name, the search for variant names is needed if we want to obtain precise data about the production of individual scientists.

By considering each author linked to his/her centre, the problem of homonyms is reduced.

A total of 474 author-centre entries were found for 307 scientists. Main differences in signing forms are due to name variants (18%) and institutional mobility (33%).

Considering the total of 474 signing forms found, 54% were automatically assigned to the studied scientists while the remaining 46% needed major revision. This result points out the importance of the latter and the limitations of using only quick search counts.

References

Bordons, M.; Zulueta, M.A.; Cabrero, A. & Barrigon, S. (1995). Identifying research teams with bibliometric tools. *Proceedings of the fifth Biennial conference of the International Society for Scientometrics and Informetrics*. (pp. 83-92) Medford: Learned information.

Ruíz Pérez, R.; Delgado López-Cózar, D. & Jiménez Contreras, E. (2002). Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies. *Journal of Medical Library Association*, 90 (4), 411-430.

Wooding, S.; Wilcox-Jay, K.; Lewison, G. & Grant, J. (2004). Co-Author Inclusion: a novel recursive algorithmic method for dealing with homonyms in bibliometric analysis. *Book of abstracts program: Eighth International Conference on Science and Technology Indicators* (p. 205). Leiden (The Netherlands): CWTS, Leiden University..