Sapientia: the Ontology of Multi-dimensional Research Assessment

Cinzia Daraio[1], Maurizio Lenzerini[1], Claudio Leporelli[1], Henk F. Moed[1], Paolo Naggar[2], Andrea Bonaccorsi[3], Alessandro Bartolucci[2]

[1] *daraio@dis.uniroma1.it; lenzerini@dis.uniroma1.it, leporelli@dis.uniroma1.it; henk.moed@uniroma1.it;*
Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, via Ariosto, 25 00185 Rome (Italy)

[2] *paolo.naggar@gmail.com; alessandro_bartolucci@fastwebnet.it*
Studiare Ltd., Rome (Italy)

[3] *a.bonaccorsi@gmail.com*
DISTEC, University of Pisa, Pisa (Italy)

Abstract
This paper proposes an Ontology-Based Data Management (OBDM) approach to a multi-dimensional research assessment. It is shown that an OBDM approach is able to take into account the recent trends in quantitative studies of Science, Technology and Innovation, including computerization of bibliometrics, multidimensionality of research assessment, altmetrics, and, more generally, the generation of new indicators with higher granularity and cross-referencing specificities according to increasingly demanding policy needs. The main features of *Sapientia* are presented, the Ontology of Multi-dimensional Research Assessment, developed within a project funded by the University of Rome La Sapienza. Illustrative examples are given of its usefulness for the specification of well known as well as recently developed indicators of research assessment.

Conference Topics
Methods and techniques; Indicators; Science policy and research assessment

Introduction: An Ontology-Based-Data-Management Approach to Multi-Dimensional Research Assessment

The quantitative analysis of Science and Technology is becoming a "big data" science, with an increasing level of "computerization", in which large and heterogeneous datasets on various aspects of Science, Technology and Innovation (STI) are combined. Within this framework, optimistic views, supporting "the end of theory" in favour of data-driven science (Kitchin, 2014), have been opposed to more critical positions in favour of theory-driven scientific discoveries (Frické, 2014) while a more balanced view emerged from a critical analysis of the current existing literature (Ekbia et al., 2015), leading the information systems community to further deeply analyse the critical challenges posed by the big data development (Agarwal, 2014). It has been rightly highlighted that "Data are not simply addenda or second-order artifacts; rather, they are the heart of much of the narrative literature, the protean stuff that allows for inference, interpretation, theory building, innovation, and invention" (Cronin, 2013, p. 435). Moreover, the need for accountability of STI activities to sustain their funding in the current difficult economic and financial situation is increasingly asking for rigorous empirical evidence to support informed policy making. Indeed, the needs to overcome the logic of rankings and the new trends in indicators development, including granularity and cross-referencing, can be explored and exploited in open data platforms with a clear description of the main concepts of the domain (Daraio & Bonaccorsi, 2015). The multidimensionality of research assessment and scholarly impact (Moed & Halevi, 2015), and the recent altmetrics movements (Cronin & Sugimoto, 2014), are questioning the traditional approach in indicators development.

Research assessment, indeed, is becoming increasingly complex due to its multi-dimensionality nature. A Report published in 2010 by the Expert Group on the Assessment of University-Based Research, installed by the European Commission proposed "a consolidated multidimensional methodological approach addressing the various user needs, interests and purposes, and identifying data and indicator requirements" (AUBR, 2010, p. 10). A key notion holds that "indicators designed to meet a particular objective or inform one target group may not be adequate for other purposes or target groups". Diverse institutional missions, and different policy environments and objectives require different assessment processes and indicators. In addition, the range of people and organizations requiring information about university-based research is growing. Each group has specific but also overlapping requirements (AUBR, 2010, p. 51).

Table 1. Main types of research outputs.

Printed outputs (texts)	**Non-printed outputs (non-text)**	**Main type of impact**
Scientific journal paper; book chapter; scholarly monograph	Research data file; video of experiment; software	Scientific-scholarly
Patent; commissioned research report;	New product or process; material; device; design; image; spin off	Economic or technological
Professional guidelines; newspaper article; communication submitted to social media, including blogs, tweets.	Interview; event; art performance; exhibit; artwork; scientific-scholarly advise;	Social or cultural

A research assessment has to take into account a range of different types of research output and impact. As regards output forms, one important distinction is between text-based and non-text based output forms. The main types are presented in Table 1. This table is not fully comprehensive. The specifications of the Panel Criteria in the Research Excellence Framework in the UK (REF, 2012, page 51 a.f.) provide more detailed lists of possible output forms arranged by major research discipline. Table 1 includes forms that are becoming increasingly important such as research data files, and communications submitted to social media and scholarly blogs. A framework for the assessment of these forms is being developed in the field of altmetrics (e.g., Taylor, 2013). The last column indicates the main types of impact a particular output may have. A distinction is made between scientific-scholarly impact, and wider impact outside the domain of science and scholarship, denoted as "societal", a concept that embraces technological, economic, social and cultural impact. A comprehensive overview of the types of impact, and the most frequently used impact indicators is presented in Table 2. The reader is referred to AUBR (2010 and Moed & Halevi (2015) for a further discussion of this table.

It is also important to include the inputs in the analysis; they should be jointly analysed with the outputs to assess the overall impact of the process (see e.g. Daraio et al., 2014, for a conditional multidimensional approach to rank higher education institutions). To meet all these new trends and policy needs a shift in the paradigm of the data integration for research assessment is needed. In this paper we advocate an OBDM approach to research assessment. This new approach radically changes the traditional paradigm of construction of STI indicators and offers a flexible and powerful tool for designing new indicators and develop rigorous policy making. The confidence in this new approach comes from three directions: (i) recent efforts from policy makers to support the creation of new datasets on S&T; (ii) bottom up standardization initiatives; (iii) development of almetrics and web-based indicators. To start with, in the last few years, several initiatives at European level have been based on an intense production and use of new data.

Table 2. Types of Research Impact and Indicators.

Type of impact	Short Description; Typical examples	Indicators (examples)
Scientific-scholarly or academic		
Knowledge growth	Contribution to scientific-scholarly progress: creation of new scientific knowledge	Indicators based on publications and citations in peer-reviewed journals and books
Research networks	Integration in (inter)national scientific-scholarly networks and research teams	(inter)national collaborations including co-authorships; participation in emerging topics
Publication outlets	Effectiveness of publication strategies; visibility and quality of used publication outlets	Journal impact factors and other journal metrics; diversity of used outlets;
Societal		
Social	Stimulating new approaches to social issues; informing public debate and improve policy-making; informing practitioners and improving professional practices; providing external users with useful knowledge; Improving people's health and quality of life; Improvements in environment and lifestyle;	▪ Citations in medical guidelines or policy documents to research articles ▪ Funding received from end-users ▪ End-user esteem (e.g., appointments in (inter)national organizations, advisory committees) ▪ Juried selection of artworks for exhibitions ▪ Mentions of research work in social media
Technological	Creation of new technologies (products and services) or enhancement of existing ones based on scientific research	Citations in patents to the scientific literature (journal articles)
Economic	Improved productivity; adding to economic growth and wealth creation; enhancing the skills base; increased innovation capability and global competitiveness; uptake of recycling techniques;	▪ Revenues created from the commercialization of research generated intellectual property (IP) ▪ Number patents, licenses, spin-offs ▪ Number of PhD and equivalent research doctorates ▪ Employability of PhD graduates
Cultural	Supporting greater understanding of where we have come from, and who and what we are; bringing new ideas and new modes of experience to the nation.	▪ Media (e.g. TV) performances ▪ Essays on scientific achievements in newspapers and weeklies ▪ Mentions of research work in social media

Legend to Table 2: Partly based on AUBR (2010) and Moed & Halevi (2015)

In the field of data on universities, the pioneering efforts of Aquameth (Daraio et al., 2011; Bonaccorsi & Daraio, 2007) and subsequently of Eumida (Bonaccorsi, 2014) have been transformed in an institutional initiative called ETER (European Tertiary Education Register), which will make publicly available microdata on universities in 2015. In the same field, the mapping of diversity of European institutions (Huisman, Meek & Wood, 2007; van Vught, 2009) led to the experimental project U-Map, after which there has been an institutional effort towards a multidimensional ranking exercise, called U-Multiranking (van Vught & Westerheijden, 2010). In the field of Public Research Organisations, there has been an effort to build up a comprehensive list of institutions and to survey their activities within the European Research Area (ERA) context. The results of the large ERA surveys, run in 2013 and 2014, will be made available in 2015. These efforts from Europe have a major counterpart on the other side of the Atlantic, where the STAR Metrics initiative (see https://www.starmetrics.nih.gov/) has promoted a federal and research institution collaboration to create a repository of data and tools that is producing extremely interesting results. All these efforts, however, are based on the construction of new datasets, or the integration of existing datasets into new ones. They do not solve the issue of comparability

and standardization of information and of inter-operability, updating and scalability of databases. It is interesting to observe that, in parallel to these efforts put in place by public institutions and policy makers, there have also been massive bottom up efforts aimed at standardizing the elementary pieces of information. Moreover, these efforts have been based on the construction of partial ontologies. Consider the following.

- ORCID (http://orcid.org/) is a non-profit organization, supported by research organizations, agencies, providers of publication management systems, and publishers, aiming at giving all researchers a unique identifier (ORCID_id number) and keeping it persistent over time. Established at the end of 2009, but operational since end 2012, it has almost reached one million researchers worldwide. Most of the increase has been achieved in a very short time frame: from 100,000 in March 2013 to almost 970,000 as of October 2014 (with 35% from European, Middle East and Asian countries);
- CERIF is a Europe-based initiative aiming at standardizing the operations of funding agencies, with the help of a full-scale ontology of almost all research products (http://www.eurocris.org);
- CASRAI (www.casrai.org) is a Canada-US initiative for the standardization of data on research institutions and funders (also supported by a committee of Science Europe; http://www.scienceeurope.org/scientific-committees/Life-sciences/life-sciences-committee);
- ISNI (www.isni.org) provides lists and metadata on higher education, research, funding and many other types of organizations, while Ringgold (www.ringgold.com) does the same in the world of publishers and intermediaries.

These initiatives are strongly supported by international scientific associations (see for example CODATA http://www.codata.org and the VIVO network of scientists: http://www.vivoweb.org/).

Finally, the rapid growth of alternative metrics and web-based metrics has also created a large space for the production of data from publicly available and other sources (Cronin & Sugimoto, 2014). Summing up, there are powerful trends that point to the need to change the overall philosophy of the production of S&T indicators. Instead of an environment in which indicators are produced in close circles, by constructing ad hoc databases, with no built-in interoperability, updating and scalability features, we have to move towards an environment in which elementary pieces of information are fully standardized, micro-data consistent with standardized definitions are (mostly) publicly available, and indicators are constructed following the policy demands on the basis of stable platforms constantly integrated and updated, instead of starting from scratch each time a new indicator is needed.

Main advantages of an OBDM approach compared to conventional data-base integration approaches

While the amount of data stored in current information systems and the processes making use of such data continuously grow, turning these data into information, and governing both data and processes are still tremendously challenging tasks for Information Technology. The problem is complicated due to the proliferation of data sources and services both within a single organization, and in cooperating environments. The following factors explain why such a proliferation constitutes a major problem with respect to the goal of carrying out effective data governance tasks:

- Although the initial design of a collection of data sources and services might be adequate, corrective maintenance actions tend to re-shape them into a form that often diverges from the original conceptual structure.
- It is common practice to change a data source (e.g., a database) so as to adapt it both to specific application-dependent needs, and to new requirements. The result is that

data sources often become data structures coupled to a specific application (or, a class of applications), rather than application-independent databases.
- The data stored in different sources and the processes operating over them tend to be redundant, and mutually inconsistent, mainly because of the lack of central, coherent and unified coordination of data management tasks.

The result is that information systems of medium and large organizations are typically structured according to a "sylos"-based architecture, constituted by several, independent, and distributed data sources, each one serving a specific application. This poses great difficulties with respect to the goal of accessing data in a unified and coherent way. Analogously, processes relevant to the organizations are often hidden in software applications, and a formal, up-to-date description of what they do on the data and how they are related with other processes is often missing. The introduction of service-oriented architectures is not a solution to this problem per se, because the fact that data and processes are packed into services is not sufficient for making the meaning of data and processes explicit. Indeed, services become other artifacts to document and maintain, adding complexity to the governance problem. Analogously, data warehousing techniques and the separation they advocate between the management of data for the operation level, and data for the decision level, do not provide solutions to this challenge. On the contrary, they also add complexity to the system, by replicating data in different layers of the system, and introducing synchronization processes across layers. All the above observations show that a unified access to data and an effective governance of processes and services are extremely difficult goals to achieve in modern information systems. Yet, both are crucial objectives for getting useful information out of the information system, as well as for taking decisions based on them. This explains why organizations spend a great deal of time and money for the understanding, the governance, the curation, and the integration of data stored in different sources, and of the processes/services that operate on them, and why this problem is often cited as a key and costly Information Technology challenge faced by medium and large organizations today (Bernstein & Haas, 2008).

We argue that ontology-based data management (OBDM, Lenzerini, 2011) is a promising direction for addressing the above challenges. The key idea of OBDM is to resort to a three-level architecture, constituted by the ontology, the sources, and the mapping between the two. The ontology is a conceptual, formal description of the domain of interest to a given organization (or, a community of users), expressed in terms of relevant concepts, attributes of concepts, relationships between concepts, and logical assertions characterizing the domain knowledge. The data sources are the repositories accessible by the organization where data concerning the domain are stored. In the general case, such repositories are numerous, heterogeneous, each one managed and maintained independently from the others. The mapping is a precise specification of the correspondence between the data contained in the data sources and the elements of the ontology.

The main purpose of an OBDM system is to allow information consumers to query the data using the elements in the ontology as predicates. In this sense, OBDM can be seen as a form of information integration, where the usual global schema is replaced by the conceptual model of the application domain, formulated as an ontology expressed in a logic-based language. With this approach, the integrated view that the system provides to information consumers is not merely a data structure accommodating the various data at the sources, but a semantically rich description of the relevant concepts in the domain of interest, as well as the relationships between such concepts. The distinction between the ontology and the data sources reflects the separation between the conceptual level, the one presented to the client, and the logical/physical level of the information system, the one stored in the sources, with the

mapping acting as the reconciling structure between the two levels. This separation brings several potential advantages:

- The ontology layer in the architecture is the obvious mean for pursuing a declarative approach to information integration, and, more generally, to data governance. By making the representation of the domain explicit, we gain re-usability of the acquired knowledge, which is not achieved when the global schema is simply a unified description of the underlying data sources.
- The mapping layer explicitly specifies the relationships between the domain concepts on the one hand and the data sources on the other hand. Such a mapping is not only used for the operation of the information system, but also for documentation purposes. The importance of this aspect clearly emerges when looking at large organisations where the information about data is widespread into separate pieces of documentation that are often difficult to access and rarely conforming to common standards. The ontology and the corresponding mappings to the data sources provide a common ground for the documentation of all the data in the organisation, with obvious advantages for the governance and the management of the information system.
- A third advantage has to do with the extensibility of the system. One criticism that is often raised to data integration is that it requires merging and integrating the source data in advance, and this merging process can be very costly. However, the ontology-based approach we advocate does not impose to fully integrate the data sources at once. Rather, after building even a rough skeleton of the domain model, one can incrementally add new data sources or new elements therein, when they become available, or when needed, thus amortising the cost of integration. Therefore, the overall design can be regarded as the incremental process of understanding and representing the domain, the available data sources, and the relationships between them. The goal is to support the evolution of both the ontology and the mappings in such a way that the system continues to operate while evolving, along the lines of "pay-as-you-go" data integration pursed in the research on data-spaces (Sarma et al., 2008).

The notions of ODBM were introduced in (Calvanese et al. 2007; Poggi et al. 2008), and originated from several disciplines, in particular, Information Integration, Knowledge Representation and Reasoning, and Incomplete and Deductive Databases. The central notion of OBDM is therefore the ontology, and reasoning over the ontology is at the basis of all the tasks that an OBDM system has to carry out. In particular, the axioms of the ontology allow one to derive new facts from the source data, and these inferred facts greatly influence the set of answers that the system should compute during query processing. In the last decades, research on ontology languages and ontology inferencing has been very active in the area of Knowledge Representation and Reasoning. Description Logics (DLs, Baader et al., 2007) are widely recognized as appropriate logics for expressing ontologies, and are at the basis of the W3C standard ontology language OWL. These logics permit the specification of a domain by providing the definition of classes and by structuring the knowledge about the classes using a rich set of logical operators. They are decidable fragments of mathematical logic, resulting from extensive investigations on the trade-off between expressive power of Knowledge Representation languages, and computational complexity of reasoning tasks. Indeed, the constructs appearing in the DLs used in OBDI are carefully chosen taking into account such a trade-off (Calvanese et al., 2007).

As indicated above, the axioms in the ontology can be seen as semantic rules that are used to complete the knowledge given by the raw facts determined by the data in the sources. In this sense, the source data of an OBDI system can be seen as an incomplete database, and query answering can be seen as the process of computing the answers logically deriving from the

combination of such incomplete knowledge and the ontology axioms. Therefore, at least conceptually, there is a connection between OBDM and the two areas of incomplete information (Imielinski & Lipski, 1984) and deductive databases (Ceri et al., 1990).

Sapientia at a glance

The main objective of *Sapientia* is to model all the activities relevant for the evaluation of research and for assessing its impact. For impact, in a broad sense, we mean any effect, change or benefit, to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia.

Table 3. Modules of the Sapientia Ontology.

N.	Module Name	Module Description
1	Overview	presents the terminological inventory needed to define the ontology domain: what is to be known to assess research activities and their impact on human knowledge and the economic system
2	Agent	models the individuals involved in the world of research, carrying out knowledge-related activities
3	Activity	models the main knowledge related activities matching them with public and relevant commitments of the agents involved in the domain (each module from 4 to 11 is devoted to a kind of knowledge-related activity - the module name corresponds to the appropriate specialization of the concept *Activity*)
4	*Research activity*	models, among the knowledge-related activities, those that allow the scientific community to advance the state of the art of knowledge
5	*Educational_activity*	models, among the knowledge-related activities, those that allow people to improve their knowledge
6	*Conferring degrees activity*	models, among the knowledge-related activities, those that grant degrees allowing people to widely qualify themselves
7	*Publishing activity*	models, among the knowledge-related activities, those that allow people to know the results of research activities
8	*Preservation activity*	models, among knowledge-related activities, those that permit the preservation of the value of things (related to research activities)
9	*Funding activity*	models, among the knowledge-related activities, those that assign and distribute the funds needed to carry out research, educational and service activities
10	*Inspecting activity*	models, among the knowledge-related activities, those that control and assess research, educational and service activities
11	*Producing activity*	models, among the knowledge-related activities, those that produce economic, society and cultural value
12	Space	models the space and its roles
13	Taxonomy	models the relevant taxonomies that classify the elements of the domain
14	Time	models the depth of time of the domain (this module is spread through the others)

Hence, *Sapientia* covers what is to be known about assess research activities and their impact on human knowledge and the economic system. For this purpose the ontology embraces:

- the inter-relationships between research activities (Modules Research_activity, Publishing_activity);

• the relationships between research activities and people's personal knowledge (Modules Teaching_activity, Conferring_degrees_activity, Publishing_activity, Producing_activity);
• the relationships between research activities and other missions of individuals and institutions (Modules Inspect_activity, Producing_activity);
• the relationship between research activities and the knowledge locally available to the companies in the economic system, enabling their innovative behavior (Module Producing_activity).

The *Sapientia* ontology includes also the activities that are needed for fostering these relationships (Modules Preservation_activity, Inspecting_activity and Funding_activities). The 14 modules that compose *Sapientia* are listed in Table 3.

Modelling choices

We pursued a modelling approach based on processes, which were conceived as collections of activities. A process is composed by inputs and outputs. Individuals and activities are the main pillars of the ontology.

We consider the building of descriptive, interpretative, and policy models of our domain as a distinct step with respect to the building of the domain ontology. However, the ontology will intermediate the use of data in the modelling step, and should be rich enough to allow the analyst the freedom to define any model she considers useful to pursue her analytic goal.

Obviously, the actual availability of relevant data will constrain both the mapping of data sources on the ontology, and the actual computation of model variables and indicators of the conceptual model. However, the analyst should not refrain from proposing the models that she considers the best suited for her purposes, and to express, using the ontology, the quality requirements, the logical, and the functional specification for her ideal model variables and indicators. This approach has many merits, and in particular:

- it allows the use of a common and stable ontology as a platform for different models;
- it addresses the efforts to enrich data sources, and verify their quality;
- it makes transparent and traceable the process of approximation of variables and models when the available data are less than ideal;
- it makes use of every source at the best level of aggregation, usually the atomic one.

More generally, this approach is consistent with the effort of avoiding "the harm caused by the blind symbolism that generally characterizes a hasty mathematization" put forward by Georgescu Roegen in his seminal work on production models and on methods in economic science (Georgescu-Roegen, 1970, 1971, 1979). In fact, one can verify the logical consistency of the ontology and compute answers to unambiguous logical queries.

Moreover, the proposed ontology allows us to follow the Georgescu-Roegen approach also in the use of the concept of process. We can analyze the knowledge production activities, at an atomic level, considering their *time* dimension and such *funds* as the cumulated results of previous research activities, both those available in relevant publications, and those embodied in the authors' competences and potential, the infrastructure assets, and the time devoted by the group of authors to current research projects. Similarly, we can analyze the output of teaching activities, considering the joint effect of *funds* such as the competence of teachers, the skills and the initial education of students, and educational infrastructures and resources. Thirdly, service activities of research and teaching institutions provide infrastructural and knowledge assets that act as a *fund* in the assessment of the impact of those institutions on the innovation of the economic system. The perimeter of our domain should allow us to consider the different channels of transmission of that impact: mobility of researchers, career of alumni, applied research contracts, joint use of infrastructures, and so on. In this context,

different theories and models of the system of knowledge production could be developed and tested (Etzkowitz & Leydesdorff, 2000).

Table 4. Indicators considered for the test of the completeness of Sapientia.

#	Indicator (I)	*Sapientia's Modules*												
		2	3	4	5	6	7	8	9	10	11	12	13	14
I1	Number of published articles	A					F					F,D1	D1	D2
I2	Number of citations	A					F					F,D1	D1	D2
I3	Citations per article	A					F					F,D1	D1	D2
I4	Normalized citation rate	A					F					F,D1	D1	D2
I5	Highly cited publications	A					F					D1	D1	D2
I6	Journal Impact Factor	A					F						F	D2
I7	Subject Normalized Impact Factor	A					F						F	D2
I8	Scimago Journal Ranking Impact Factor	A					F						F	D2
I9	H-index	F					F					A	F	D2
I10	E-index	F					F					A	F	D2
I11	Number of patents	A									F	F,D1	D1	D2
I12	Full text article download count	A					F					F,D1	D1	D2
I13	Mentions in social media	A					F					F,D1	D1	D2
I14	Research output per academic staff	A					F					F,D1	D1	D2
I15	Percentage of Highly Cited Publications	A					F					D1	D1	D2
I16	Number of keynote addresses at conferences	A					F					F,D1	D1	D2
I17	Number of prestigious awards and prizes	A								F		F,D1	D1	D2
I18	Number of visiting research appointments	F,A										D1	D1	D2
I19	Member of editorial board	A					F					D1	D1	D2
I20	Refereeing activity for journals	A	F							F		D1	D1	D2
I21	External research income	A							F		F	D1	D1	D2
I22	Number of competitive grants won	A							F		F	D1	D1	D2
I23	Percentage of competitive grants	A							F		F	D1	D1	D2
I24	External research income per academic staff	A							F		F	F,D1	D1	D2
I25	Employability of PhD graduates	A				F						D1	F,D1	D2
I26	Commerc. of research generated intellectual property	A									F	D1	D1	D2
I27	End-users esteem	A								F		D1	D1	D2
I28	Number of funding from end-users	A,D1							F			D1	D1	D2
I29	Percentage of funding from end-users	A,D1							F			D1	D1	D2
I30	Post-graduate research student load	A		F	F							D1	D1	D2
I31	Involvement of early career researchers in teams	A,F		F								D1	D1	D2
I32	Number of collaborations and partnerships	F		A								F,D1	D1	D2
I33	Doctoral completions	A				F						D1	D1	D2
I34	Research active academics	F,A					F					D1	D1	F,D2
I35	Percentage of research active per total academic staff	F,A					F					D1	D1	F,D2
I36	Total R&D investment	A							F			D1	D1	D2
I37	Research Infrastructures and facilities	A					F	F				D1	D1	D2
I38	Research ethics						F					A,F	F,D1	D2

Testing the Ontology: analysis of the competency questions

One way to check if the ontology contains all the relevant information and/or details to represent the domain of interest, currently used in knowledge representation, is based on the specification of competency questions (Gruninger & Fox 1995). These questions correspond to check whether the ontology contains enough information to answer these types of questions or whether the answers require a particular level of detail or representation of a particular module of the ontology that needs to be further developed. The analysis of the competency questions of *Sapientia* has been carried out on the indicators contained in the paper by Moed and Halevi (2015), integrated with the additional indicators reported in the AUBR (2010) document. In addition, other key references of the ontological commitments have been Moed, Glanzel and Schmock (2004), Moed (2005) and Cronin and Sugimoto (2014), together with the knowledge background of the team of the project.

Table 4 contains the list of indicators considered for the verification of the competency questions. Associated to each indicator are reported the following pieces of information:

- Facts (F) are the content of the data, the relevant information about atomic events relevant for the construction of the indicator;
- Aggregation level (A) is the minimal aggregation level: the concept which classifies the objects included in the indicator;
- Dimensions of the analysis (D), are descriptive properties which are relevant to access higher level of aggregation. They are evaluated by the dimension of taxonomy (D1) and that of time (D2).

Table 5 summarizes the number of facts (F), aggregations (A) and dimensions (D) by module, as reported in Table 4, to check the comprehensiveness of *Sapientia* with respect to the indicators listed therein. Put it in another way, we checked whether our ontology was able to include all the relevant conceptual information requested by the specification of the listed indicators in Table 4. The answer to this question is indeed positive.

Table 5. Some statistics on the "usage" of the Ontology modules.

	2	3	4	5	6	7	8	9	10	11	12	13	14
F	7	1	2	1	2	20	1	7	3	6	13	5	2
A	34	0	1	0	0	0	0	0	0	0	2	1	0
D	2	0	0	0	0	0	0	0	0	0	31	31	38

By inspecting Table 5 it clearly appears that only a few modules are used for the specification of the indicators reported in Table 4. This means that our ontology covers a much broader conceptual domain with respect to the one underlying (even if not formally specified) by the indicators reported in Table 4. The most frequently used module is the Publishing module (7), followed by Space (12) and Funding (9). We note that the modules 12 (Space), 13 (Taxonomy) and 14 (Time) are used in the majority of the cases to further characterize the dimensions of the considered indicators.

A new way to conceive and specify STI indicators

By adopting an OBDM perspective a new approach to designing indicators can be implemented. This new approach aligns very well with the recent trends described in the introduction.

The traditional approach to indicators' design is based on informal definitions expressed in a natural language (English, typically). An indicator is defined as a relationship between variables, e.g. a ratio between number of publications per academic staff, chosen among a predefined set of data collected and aggregated ad hoc, by a private or a public entity, according to the user needs, and hence not re-usable for future assessment and use.

The OBDM approach we pursue in this paper permits a *more advanced specification* of an indicator according to the following dimensions:

- the *ontological dimension*. It represents the domain (portion) of the reality to be measured by the indicator (obviously, in the scope of this paper, all indicators will share the *Sapientia* ontology as their ontological part);
- the *logical dimension*. It denotes the question that has to be asked to the ontological portion in order to retrieve all the information (data) needed for calculating the indicator value. In this case the data are extracted from the sources through the mapping considering the logical specification of the query;
- the *functional dimension*. It indicates the mathematical expression that has to be applied on the result of the logical extraction of data carried out in the previous point in order to calculate the indicator value;
- the *qualitative dimension*. It specifies the questions that have to be asked to the ontological part in order to generate the list of problems affecting the meaningfulness of the calculated indicator. An indicator will be considered meaningful if the list of its problems is empty.

In addition to the advantages of the OBDM recalled in previous sections above, the main specific benefits of this approach for designing indicators are the following:

1. It offers a space to *freely* explore the *generation of new indicators,* not previously specified by users, thanks to the *multiple inheritance* in the hierarchy of the concepts (a concept can be subsumed in several concepts).
2. For standard indicators specified by the users it can be seen immediately what is *missing* or which *problems* exist to calculate them;
3. It provides more alternatives and diagnostic ways to check the *robustness* of indicators with respect to opportunistic behaviour and the general goals of the assessment;
4. The formal specification of the indicators is made *independently* of the data. In this way, when applied to heterogeneous data sources, OBDM offers the opportunity to compute "comparable" indicator values at different level of aggregation. Moreover, it offers a reference system to *check the comparability* level among the heterogeneous sources of data and to identify where to invest in order to overcome the remaining existing comparability problems.
5. This approach permits an *unambiguous* way to define and compute the indicators. The indicator is calculated always in the same way.

Conclusions and further developments

In this paper we advocated the use of an OBDM approach to research assessment. We explained the reasons why a paradigm shift in research assessment is needed and outlined the main advantages of an OBDM approach over traditional databases integration approaches. We described the main objectives and structure of *Sapientia* the Ontology of Multi-dimensional Research Assessment. Finally, we illustrate the new indicator design methodology implicitly provided by an OBDM approach.

Sapientia 1.0 has been closed on the 22nd December 2014 and consisted of around 350 symbols (including concepts, relations and attributes). The full documentation of the Ontology is under way together with the mapping with several sources of data. Due to the works on the documentation and the mapping with the data in progress, as well as the limited number of pages available, we concentrated our presentation on the methodological aspects related to the development of the Sapientia.

We believe in fact that it will open a new stream of studies to further explore and exploit the OBDM approach for STI indicator designers and policy makers.

Acknowledgments

The financial support of the "Progetto di Ateneo 2013", University of Rome La Sapienza, is gratefully acknowledged.

References

Agarwal, R., & Dhar, V. (2014). Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. I*nformation Systems Research, 25*(3), 443-448.

AUBR Expert Group (2010). Expert Group on the Assessment of University-Based Research. Assessing Europe's University-Based Research. European Commission – DG Research. EUR 24187 EN

Baader F., D. Calvanese, D. McGuinness, D. Nardi, & P. F. Patel-Schneider, (eds) (2007). *The Description Logic Handbook: Theory, Implementation and Applications.* Cambridge University Press, 2nd edition.

Bernstein P. A. & Haas L.(2008). Information integration in the enterprise. *Communication of the ACM*, *51*(9), 72–79.

Bonaccorsi A., & Daraio C. (eds.) (2007) *Universities and strategic knowledge creation. Specialization and performance in Europe*. Cheltenham, Edward Elgar.

Bonaccorsi, A. (ed.) (2014) *Knowledge, diversity and performance in European higher education*. Cheltenham, Edward Elgar.

Calvanese D., G. De Giacomo, D. Lembo, M. Lenzerini, & R. Rosati (2007), Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *Journal of Automated Reasoning*, *39*(3), 385–429.

Ceri S., G. Gottlob, & L. Tanca (1990). *Logic Programming and Databases*. Springer, Berlin (Germany).

Console M., Lembo D., Santarelli V. & Savo D.F. (2014a). Graphol: Ontology Representation Through Diagrams. *Proc. of the 27th Int. Workshop on Description Logic*.

Console M., Lembo D., Santarelli V., & Savo D.F. (2014b). Graphical Representation of OWL 2 Ontologies through Graphol. *Proc. of the 13th International Semantic Web Conference Posters & Demos*.

Cronin B. & Sugimoto C. (ed) (2014). Beyond bibliometrics. Harnessing multidimensional indicators of scholarly impact. MIT Press, Cambridge Mass.

Cronin, B. (2013). Thinking about data. *Journal of the American Society for Information Science and Technology*, *64*(3), 435–436.

Daraio, C. et al. (2011). The European university landscape: A micro characterization based on evidence from the Aquameth project. *Research Policy 40*, 148–164.

Daraio C. & Bonaccorsi A. (2015). Beyond university rankings? Generating new indicators on universities by linking data in open platforms. under review for *JASIST*.

Daraio, C., Bonaccorsi A., & Simar L. (2015). Rankings and university performance: a conditional multidimensional approach. *European Journal of Operational Research*, *244*, 918–930.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., ... & Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*.

Etzkowitz H., & L. Leydesdorff. (2000). The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations, *Research Policy*, *29*(2), 109-123.

Frické, M. (2014). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, 66(4), 651-661.

Georgescu-Roegen, N. (1970). The economics of production. *The American Economic Review, 60*(2), 1-9.

Georgescu-Roegen, N. (1972). Process analysis and the neoclassical theory of production, *American Journal of Agricultural Economics,* 279-294.

Georgescu-Roegen, N. (1979). Methods in economic science, *Journal of Economic Issues,* 317-328.

Gruninger, M. & Fox, M.S. (1995). Methodology for the Design and Evaluation of Ontologies. In: *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing*, IJCAI-95, Montreal.

Huisman, J., Meek, V.L. & Wood, F.Q. (2007). Institutional diversity in higher education: a cross-national and longitudinal analysis, *Higher Education Quarterly,* 61/4: 563-577.

Imielinski T. & W. Lipski, Jr. (1984) Incomplete information in relational databases. *Journal of the ACM, 31*(4), 761–791.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 1-12.

Lenzerini M. (2011). Ontology-based data management, *CIKM* 2011, 5-6.

Moed, H.F. (2005). *Citation Analysis in Research Evaluation*, Springer NY.

Moed, W. Glanzel & U. Schmoch (ed.) (2004), *Handbook of Quantitative Science and Technology Research,* Kluwer Academic Publishers, 51-74.

Moed, H. F., & Halevi, G. (2015). The Multidimensional Assessment of Scholarly Research Impact, *Journal of the American Society for Information Science and Technology*, forthcoming.

Parnas, D. L. (1972). On the criteria to be used in decomposing systems into modules. *Communications of the ACM*, *15*(12), 1053-1058.

Poggi A. , D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, & R. Rosati (2008). Linking data to ontologies. *Journal on Data Semantics*, *10*,133–173.

REF (Research Excellence Framework) (2012). Panel Criteria and Working Methods. Retrieved January 7, 2015 from: http://www.ref.ac.uk/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.pdf.

Sarma A. D., Dong X., & Alon Y (2008). Bootstrapping pay-as-you-go data integration systems. *Proc. of ACM SIGMOD*, 861–874.

Taylor, M. (2013). Exploring the Boundaries: How Altmetrics Can Expand Our Vision of Scholarly Communication and Social Impact. *Information Standards Quarterly*, *25*, 27-32.

Van Vught, F. (ed.) (2009). Mapping the higher education landscape: Towards a European classification of higher education. Dordrecht: Kluwer.

Van Vught, F., & Westerheijden, D.F. (2010). Multidimensional ranking: a new transparency tool for higher education and research, *Higher Education Management and Policy 22*/3, 1-26.