ISSN 2281-4299

DIPARTIMENTO DI INGEGNERIA INFORMATICA
AUTOMATICA E GESTIONALE ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

Data Integration for Research and Innovation Policy: An Ontology-based Data Management Approach

Cinzia Daraio, Maurizio Lenzerini, Claudio Leporelli,
Henk F. Moed, Paolo Naggar, Andrea Bonaccorsi,
Alessandro Bartolucci

Technical Report n. 10, 2015

Data Integration for Research and Innovation Policy: An Ontology-based Data Management Approach[1]

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 *(corresponding author);* 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 (Italy)

Paper prepared for the Special Session at the STI/ENID Conference 2015, Lugano 2-4 September 2015, "Towards Standardisation, Harmonisation and Integration of Data from Heterogeneous Sources for Funding and Evaluation Purposes" organised by Wolfgang Glänzel and Hans Willems.

Abstract

The main objective of this paper is to propose an Ontology-based Data Management (OBDM) approach to coordinate, integrate and maintain the data needed for Science, Technology and Innovation (STI) policy development. The OBDM approach we propose is a form of integration of information in which the global schema of data is substituted by the conceptual model of the domain, formally specified through an ontology.
Our approach, implemented in the *Sapientia* ontology (*Sapientia*: the Ontology of Multi-Dimensional Research Assessment) offers a transparent platform on which to base the evaluation process; permits to define and specify in an unambiguous way the indicators on which the evaluation is based on; allows us to track their evolution over time; makes it possible the analysis of the feedbacks of the indicators on the behavior of scholars and allows us to find out opportunistic behaviors; provides a monitoring system to track over time the changes in the established evaluation criteria and their consequences on the research system. We claim that an higher availability and a more transparent view on the scholarly outcomes may improve the understanding of basic science from the broad society and can improve the communication of the research outcome to the public opinion, which, in the present economic phase, has an increasingly money-for-value approach about the funding of science.
A lot of work on these issues has still to be carried out. Nevertheless we believe that a new line of research based on an OBDM approach could successfully contribute to solve some of the key issues in the integration of heterogeneous data for STI policies.

[1] This work is based on two papers accepted for presentation and published in the proceedings of the ISSI 2015 Conference (see Daraio, Lenzerini et al. 2015a, b).

1. Introduction

The recent trends in research assessment, the development of altmetrics, the crucial role of data together with the complexity of research assessment, granularity and increasingly demanding policy needs call for new ways of data integration and management.
There have been several initiatives of governments and research projects on these matters. However, the main problems of integration of data on Science, Technology and Innovation, such as the data quality issues; the comparability problems; the lack of standardization, interoperability and modularization; the difficulties in the creation of concordance tables among different classification schemes; the difficult and costly extension and update of the integrated database, are far from being solved.

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 are combined. In this context, understanding and formally specifying the meaning of data is of paramount importance.
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).
The necessity of providing accountability of Science, Technology and Innovation (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.
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 complexity of the multidimensionality of research assessment and scholarly impact (Moed & Halevi 2015) is questioning the traditional approach in indicators development. 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).
The assessment of research has to take into account a range of different types of research output and impact. See Table 1 for a non-exhaustive outline: it includes forms that are becoming increasingly important such as research data files, and communications submitted to social media and scholarly blogs. The last column indicates the main types of impact a particular output may have. A distinction is made between scientific-scholarly impact, and more wider impact outside the domain of science and scholarship, denoted as "societal", a concept that embraces technological, economic, social and cultural impact.
A more detailed list of possible outputs by research area is reported in the specifications of the Panel Criteria in the Research Excellence Framework in the UK (REF 2012, page 51.). See also AUBR (2010) and Moed & Halevi (2015) for further details.
It is also important to include the inputs in the research assessment process; they should be jointly analysed with the outputs to assess the overall impact of the process (see e.g. Daraio et al. 2015, 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 data integration for research assessment is needed. In this paper we advocate an OBDM approach to integrate heterogeneous data sources, including big scholarly data (such as publications and citations) to support the assessment of research and develop "science of science" policy models.

Type of impact ***(examples of printed and non printed outputs)***	**Short Description; Typical examples**	**Indicators (examples)**
Scientific-scholarly or academic *(printed outputs: Scientific journal paper; book chapter; scholarly monograph; non-printed outputs: Research data file; video of experiment; software)*		
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;
Economic or Technological *(Printed outputs: Patent; commissioned research report. Non printed outputs: New product or process; material; device; design; image; spin off)*		
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
Societal or cultural *(printed outputs: Professional guidelines; newspaper article; communication submitted to social media, including blogs, tweets. Non printed outputs: Interview; event; art performance; exhibit; artwork; scientific-scholarly advise)*		
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
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

Table 1: Types of Research Outputs, Impacts and Indicators (Source: adapted from Moed and Halevi, 2015)

The paper unfolds as follows. In the next section we illustrate the main problems of heterogeneous data integration. Section 3 presents the main advantages of an OBDM approach and outlines its implementation through *Sapientia*, the ontology of multidimensional research assessment. Section 4 illustrates the usefulness of an OBDM approach to specify STI indicators in an innovative way. Section 5 shows how an OBDM approach may be useful to develop science of science policy models, while Section 6 concludes the paper.

2. Difficulties in accessing and managing distributed and heterogeneous data

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 management, 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).

In the next section we advocate for an Ontology-based Data Management (OBDM, Lenzerini 2011) approach as a promising direction for addressing the above challenges.

3. Our proposal: an Ontology-Based Data Management Approach (OBDM)

In this paper we argue that the Ontology of the Multi-Dimensional Research Assessment (*Sapientia*, created within a research project funded by the university of Rome La Sapienza) with its underlying OBDM approach may be a powerful tool to coordinate, integrate and maintain the data needed for Science, Technology and Innovation policy development.

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 users 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 scheme 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 users 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 user, 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.
Firstly, 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.
Secondly, 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 (Sarma et al., 2008). See Table 2 which summarizes the main advantages of the OBDM approach.

Advantage	Short Description
Conceptual access to the data	Users can access the data by using the elements of the ontology.
Re-usability	By making the representation of the domain explicit, we gain re-usability of the acquired knowledge.
Documentation and standardization	The mapping layer explicitly specify the relationships between the domain concepts and the data sources. It is useful for documentation and standardization purposes.
Flexibility of the system	You do not have to merge and integrate all the data sources at once which could be extremely costly.
Extensibility of the system	You can incrementally add new data sources or new elements (ability to follow the incremental understanding of the domain) when they become available.
Opening of the system	Provide a conceptual framework which can be used as a common language by the community.

Table 2. Main advantages of an OBDM approach over a traditional "sylos"-based approach

The notions of OBDM were introduced in Calvanese et al. (2007), Poggi et al. (2008), Lenzerini (2011), 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 OBDM 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 OBDM 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).

The OBDM approach has been implemented in a research assessment framework within a research project funded by the University of Rome La Sapienza, which produced as an output *Sapientia* the Ontology of Multidimensional research assessment[2].
The main objective of *Sapientia* (the Ontology of Multidimensional Research Assessment) 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 (REF, 2012). *Sapientia* 1.0 was closed the 22nd of December 2014, and was organized in 14 Modules (Overview, Agent, Activity, Research Activity, Educational Activity, Conferring degrees activity, Publishing activity, Preservation activity, Funding activity, Inspecting activity, Producing activity, Space, Taxonomy and Time), including around 350 symbols (concepts, relations and attributes).
We are consolidating our ontology (*Sapientia*), completing its documentation and investigating the interoperability of *Sapientia* with other existing initiatives, such as STAR Metrics, CERIF (http://www.eurocris.org) CASRAI (www.casrai.org); ISNI (www.isni.org) and so on. We found that our ontology is *complementary* with respect to the existing initiatives and the top-down approach we followed to its design and development is *fully interoperable* with existing initiatives cited above. *Sapientia* will be published on-line afterwards.
The current version of Sapientia, version 2.0, includes 11 modules that are organized according to Figure 1, whose main agents and activities for each module are reported in Figure 2.

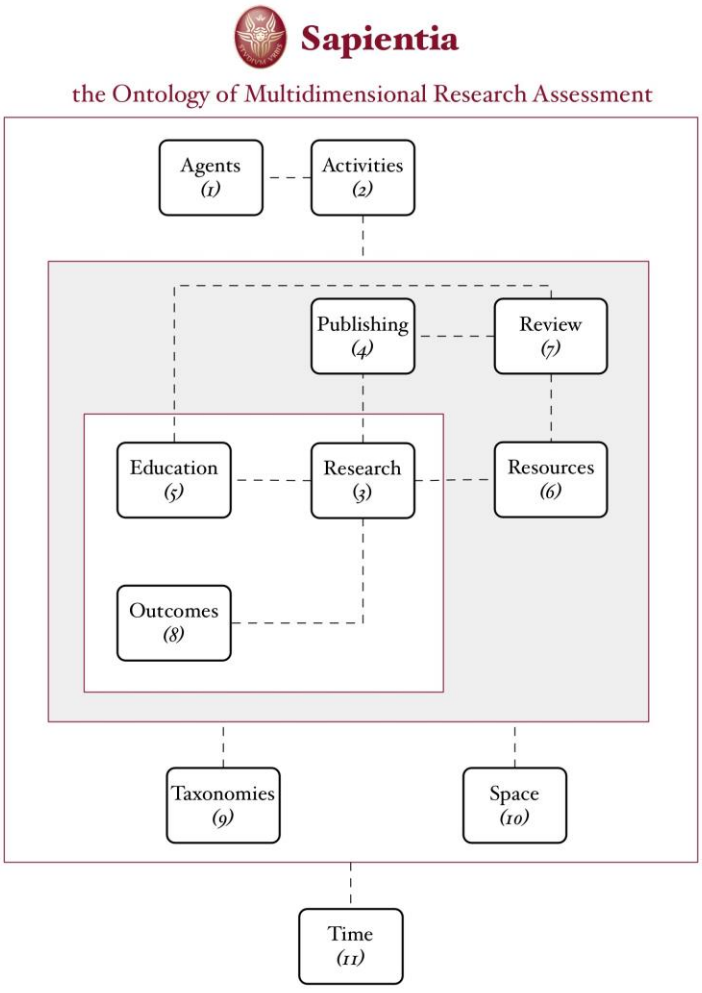

Figure 1. The 11 Modules of *Sapientia 2.0*: the Ontology of Multidimensional Research Assessment.

[2] Sapientia 1.0 has been presented at the Workshop of the 20 February 2015 held at DIAG, Sapienza University of Rome whose proceedings are reported in Daraio (2015).

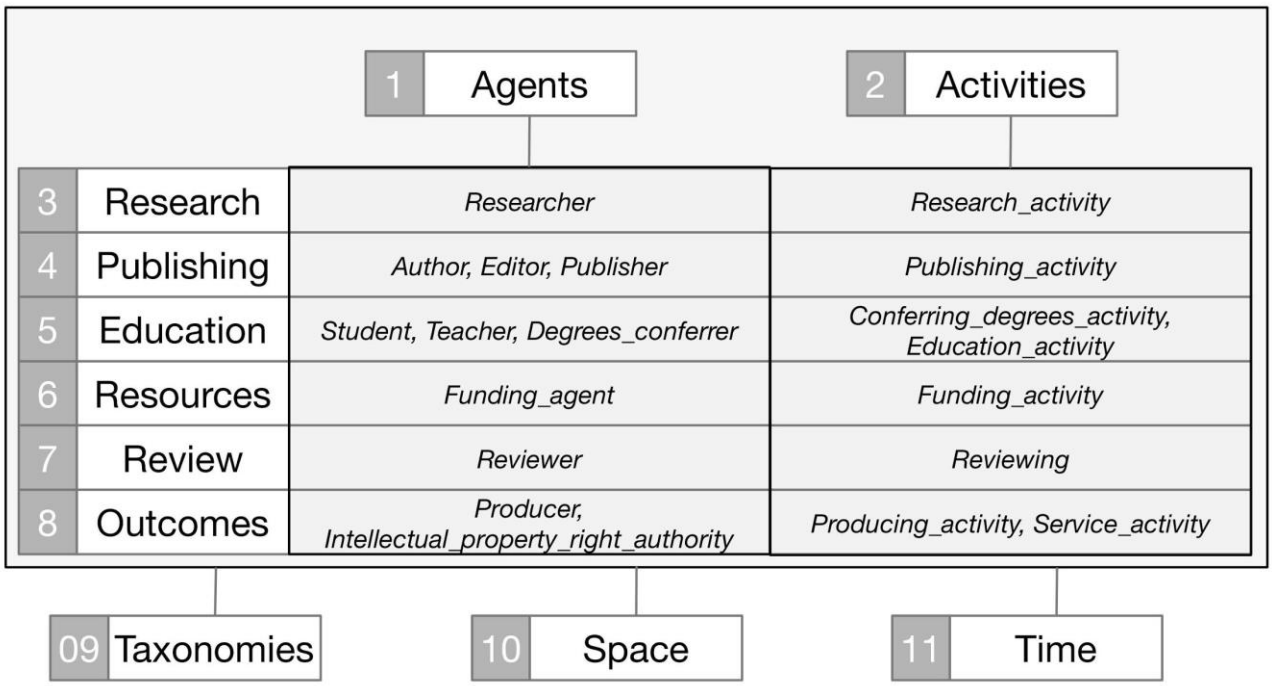

Figure 2. Main agents and activities of *Sapientia 2.0*.

As illustrated in Figure 1, the *Sapientia* ontology models the main activities (Module 2) carried out by the agents (Module 1). It includes a core set of modules which are Research (Module 3), Education (Module 4) and production, including services and other third mission activities (Module 8). These activities are part of an extended set of modules which includes an ancillary module of Research (Module 4 Publishing) and other two modules containing relevant activities to foster the relationships among the core set of modules (i.e., Modules 6 Resources, including funding and projects, and Module 7 Review). The 11 modules that compose *Sapientia* are briefly described in Table 3.

N.	Module Name	Module Description
1	**Agents**	It models the individuals involved in the broad world of research, carrying out knowledge-related activities.
2	**Activities**	It models the main knowledge related activities matching them with public and relevant commitments of the agents involved in the domain (each module from 3 to 8 is devoted to a kind of knowledge-related activity - the module name corresponds to the module appropriate specialization of the concept Activity).
3	**Research**	It models, among the knowledge-related activities, those that allow the scientific community to advance the state of the art of knowledge.

4	**Publishing**	It models, among the knowledge-related activities, those that allow people to know the results of research activities.
5	**Education**	It models, among the knowledge-related activities, those that allow people to improve their knowledge and those that grant degrees allowing people to widely qualify themselves.
6	**Resources**	It models, among the knowledge-related activities, those that assign and distribute the funds needed to carry out research, educational and service activities.
7	**Review**	It models, among the knowledge-related activities, those that control and assess research, educational and service activities.
8	**Outcomes**	It models, among the knowledge-related activities, those that produce economic, society and cultural value.
9	**Taxonomies**	It models the relevant taxonomies that classify the elements of the domain.
10	**Space**	It models the space and its roles.
11	**Time**	It models the depth of time of the domain.

Table 3. Description of the *Sapientia* 2.0's Modules.

4. An OBDM approach to specify Science, Technology and Innovation (STI) indicators in an innovative way

The increase availability of data sources, the need to combine several assessment criteria and their actual use ask for an overarching structure to overcome the main problems in STI indicator development which are listed below (and summarized in Table 4, left column):
-Concepts are not clearly defined (e.g. what is a "publication"?)
-Informal definitions can be based on everyday language
-One concept name may refer to different concepts
-Ad hoc definitions of indicators based on available datasets or specific user needs
-Indicators non re-usable in future contexts
-Database content is not fully transparent
-Aggregate indicators cannot be decomposed into smaller units.

Problems in STI indicators design	Benefits of the OBDM approach
-Ambiguity of concepts -Existence of informal (non-codified) definitions -Ambiguity of names of concepts -Ad hoc definitions of indicators -Non re-usability of Indicators -Non-transparency of the database content -Non-decomposability of aggregate indicators	- Formal specification of the indicators independently with respect to the data; - Computation of "comparable" indicators at different level of aggregation; - A reference system to check the comparability level among the heterogeneous data sources; - Unambiguous way to define and compute the indicators; - A formal framework for concepts and data sources; - Transferability to new generations of producers and users.

Table 4. Problems in STI design and benefits of an OBDM Approach

In Daraio, Lenzerini et al. (2015a) we describe in details the ability of *Sapientia* to specify the performance indicators proposed by the AUBR (2010).

An OBDM approach offers the possibility to develop indicators according to the following dimensions (see Table 5).

Dimension	Specification
Ontological	Formal representation of a domain: objects, their properties and relationships
Logical	Data extracted from sources through mapping considering a query's logical specification
Functional	Mathematical expression to be applied to the results of the logical data extraction
Qualitative	Questions addressed to the ontology for the assessment of the indicators' meaningfulness

Table 5. Dimensions of indicators in an OBDM framework

The main benefits of this approach for indicators' designers and users (summarized in Table 4, right column) are:
- The formal specification of the indicators which is made independently of the data;
- The opportunity to compute "comparable" indicator values at different level of aggregation;
- It offers a reference system to check the comparability level among the heterogeneous data sources;
- It permits an unambiguous way to define and compute the indicators;
- The knowledge on the indicator system (concepts and data sources) is embedded in a formal framework;
- This knowledge can be transferred more easily to new generations of producers and users.

5. Using *Sapientia* for Science of Science policy

Our perspective, based on an OBDM approach, allows us to contribute to enriching the methodologies available for science of science policy (Fealing et al. 2011) and research assessment.

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. 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 permits the use of a common and stable ontology as a platform for building different models and indicators;
- 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 (see examples in the following), allowing subsequent, multilevel and multidimensional aggregations.

In this framework, exploratory data analysis, and the building of synthetic indicators, are only an intermediate step of the modelling effort that aims to the interpretation of behaviours, the explanation of differences in performance, the identification of causal chains of phenomena. That leads to the development of a policy-design model, whose inputs are policy instruments, and whose outputs are performance indicators for research activities and economic welfare.
The learning and theory building process requires feedbacks that could also concern the ontology level: the addition of new concepts and data, through the specialization of general concepts or the enlargement of the ontology commitment, could reflect the intermediate achievements of the learning process such as the necessity of improvement of the theories submitted to test.
More often, however, a well-conceived ontology will resist to the competency test implied by new model and theories, and the most serious constraint to model development will be the impossibility of a complete mapping between the ontology and the sources, i.e. the lack of data. This is a negative result only for the short-term. In the medium and long term, the dialogue within the community of researchers that use the ontology as a workbench will result in a joint effort towards other stakeholders in order to improve detail, quality, and scope of data collection.
Moreover, the shared use of logically sound definition for indicators increase the ability of the analysts to compare their studies and to test old and new theories.
Consider as an example the important issue of the assessment of the effects of scale economies on the performance of a research institution and of its affiliates. The results can widely differ if you set the analysis at different levels of aggregation: all the public research and education institutions of single countries, single universities, faculties, let's say, of Science and Technology, departments of Computer Science, research groups, or individuals within these groups.
Moreover, at different aggregation levels, the possible moderating variables or causes of different performances can widely differ. Legislation and regulation, public funding, teaching fees and duties matter at national level. Geography, characteristics of the local economic and cultural system, effectiveness of research and recruiting strategy, budgeting, infrastructures matter at the university or department level. Intellectual ability of researchers, history and stability of the group, ability to recruit doctoral students, worldwide network of contacts matter at the research groups and individuals level.

Time is a crucial dimension of research modelling. We pursue a modelling approach based on processes, i.e. collections of activities performed by agents through time, following Georgescu Roegen (1970, 1972, 979). Therefore, to represent the knowledge production activities, at an atomic level, we aim to consider both stock inputs such as the cumulated results of previous research activities (those available in relevant publications, and those embodied in the authors' competences and potential), the infrastructure assets, and flow inputs as the time devoted by the group of authors to current research projects. Similarly, we aim to analyse the output of teaching activities, considering the joint effect of resources 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 have an impact on the innovation of the economic system; therefore, 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.

To bridge the gaps existing in the literature, and to integrate existing bottom-up initiatives in a coherent theoretical-based platform, we suggest an OBDM approach.
We need a change in the overall approach to the assessment of science and technology: metrics and indicators can have negative effects on the scientific community because they encourage a reductionist philosophy; on the contrary, we propose using well-defined concepts and data to build interpretative models, in order to compare and discuss theories[3]. That can be useful both to promote a pluralistic community of analysts, and to build consensus on less superficial evaluation procedures of researchers and institutions[4]. Moreover, indicators are often produced in closed circles, collecting ad hoc databases, with no built-in interoperability, updating and scalability features.

We have to move towards an environment in which data are publicly available, collected and maintained on stable platforms, where ontologies give confidence on the precise meaning of data to people that propose models and to those that evaluate them. These repositories of knowledge can evolve following the analytical needs of the research community and the policy institutions, instead of starting from scratch each time a new research project starts. We propose our Sapientia ontology as a starting point to be opened, shared with the community and further developed and integrated with existing bottom-up initiatives as well as with new theories and paradigms.

6. Conclusions

The rapid expansion of big data and open data; the altmetrics movement; the complexity of research assessment and the more and more demanding policy needs ask for new ways of data integration and interoperability among many heterogeneous data sources, including Big Scholarly Data, such as publications and citations.
Although there have been several initiatives of governments and research projects, the main problems of integration of data on Science, Technology and Innovation are far from being solved. The existing initiatives, indeed, do not solve the main problems related to the integration of heterogeneous sources of data, such as the data quality issues; the comparability problems; the lack of standardization, interoperability and modularization; the difficulties in the creation of

[3] An interesting comparison is possible with the standard setting process in the accounting community (IFRS, 2015) and the development of taxonomies and formal languages like XBRL to communicate and manipulate accounting documents (IFRS, 2014).
[4] Even the assessment of R&D performance in a profit oriented organization will gain in insight and generality if multiple approaches (qualitative and quantitative, micro and macro) are parallel pursued and compared (Werner and Souder, 1997; Nudurupati et al., 2011).

concordance tables among different classification schemes; the difficult and costly extension and update of the integrated database built on independent and heterogeneous databases.
In this paper we argue that the Ontology of the Multi-Dimensional Research Assessment (*Sapientia*) with its underlying Ontology-based Data Management (OBDM) approach may be a powerful tool to coordinate, integrate and maintain the data needed for Science, Technology and Innovation policy development. The OBDM approach we propose is a form of integration of information in which the global schema of data is substituted by the conceptual model of the domain, formally specified through an ontology.
Our approach, implemented in the Sapientia ontology, offers a transparent platform on which to base the evaluation process; permits to define and specify in an unambiguous way the indicators on which the evaluation is based on; allows us to track their evolution over time; makes it possible the analysis of the feedbacks of the indicators on the behavior of scholars and allows us to find out opportunistic behaviors; provides a monitoring system to track over time the changes in the established evaluation criteria and their consequences on the research system. We claim that an higher availability and a more transparent views on the scholarly outcomes may improve the understanding of basic science from the broad society and can improve the communication of the research outcome to the public opinion, which, in the present economic phase, has an increasingly money-for-value approach about the funding of science.
Furthermore, our approach, by providing a stable but flexible and extensible platform, might be able to foster the involvement and contribution of scholars to the evaluation process and therefore will contribute to the development of the Web of Scholars.
Despite the fact that still a lot of research on this issue has to be carried out, we argue that this approach could be very promising for the resolution of important open questions that we have mentioned in this work and that a new line of research based on a OBDM approach could successfully contribute to solve some of the key issues raised in this paper.

Acknowledgments

Research support from the Progetto di Ateneo 2013 (C26A13ZXRY) of the Sapienza university of Rome is gratefully acknowledged.

References

Agarwal, R., & Dhar, V. (2014). Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. Information 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. Comm. of the ACM, 51(9):72–79.

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. J. of Automated Reasoning, 39(3):385–429.

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

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

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

Daraio C. (2015), (Eds.), Efficiency, Effectiveness and Impact of Research and Innovation, Proceedings of the Workshop of the 20 February 2015 DIAG, Sapienza University of Rome, Efesto Edizioni, Rome, ISBN 9788899104306.

Daraio C., & Bonaccorsi A. (2015), Beyond university rankings? Generating new indicators on universities by linking data in open platforms, Journal of the American Society for Information Science and Technology *forthcoming*.

Daraio, C., Bonaccorsi A., & Simar L. (2015), Rankings and University Performance: a Conditional Multidimensional Approach, European Journal of Operational Research, 244, 918–930.

Daraio C., Lenzerini M., Leporelli C., Moed H.F., Naggar P., Bonaccorsi A. & Bartolucci A. (2015a). Sapientia the Ontology of Multi-Dimensional Research Assessment, in Salah, A.A., Y. Tonta, A.A. Akdag Salah, C. Sugimoto, U. Al (Eds.), Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015, Bogaziçi University Printhouse, pp. 965-977.

Daraio C., Lenzerini M., Leporelli C., Moed H.F., Naggar P., Bonaccorsi A. & Bartolucci A. (2015b). Connecting Big Scholarly Data with Science of Science Policy: an Ontology-Based-Data-Management (OBDM) Approach, in Salah, A.A., Y. Tonta, A.A. Akdag Salah, C. Sugimoto, U. Al (Eds.), Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015, Bogaziçi University Printhouse, pp. 1232-1233.

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.

Fealing K. H., Lane J. I., Marburger J. H. JIII, & Shipp S. S. (Eds.) (2011), The Science of Science Policy, A Handbook. Stanford, USA, Stanford University Press.

Frické, M. (2014). Big data and its epistemology. Journal of the Association for Information Science and Technology.

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

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

Georgescu-Roegen, N. (1979) Methods in economic science, Journal of economic issues (1979): 317-328.

IFRS (2014) A guide to understanding IFRS Taxonomy update. IFRS Taxonomy Guides.

IFRS (2015) Conceptual Framework for Financial Reporting. Exposure Draft ED/2015/3

Imielinski T. & W. Lipski, Jr.(1984) Incomplete information in relational databases. J. 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., & Halevi, G. (2015). The Multidimensional Assessment of Scholarly Research Impact, Journal of the American Society for Information Science and Technology, 66 (10), 1988–2002.

Nudurupati, S.S., U.S. Bititci, V. Kumar, F.T.S. Chan. State of the art literature review on performance measurement. Computers & Industrial Engineering 60 (2011) 279–290

Poggi A., D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, & R. Rosati. (2008). Linking data to ontologies. J. on Data Semantics, X: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), Halevy. Bootstrapping pay-as-you-go data integration systems. In Proc. of ACM SIGMOD 2008, pages 861–874.

Werner, B. M., & Souder, W. E. (1997). Measuring R&D performance--state of the art. Research technology management, 40(2), 34.