Connecting Big Scholarly Data with Science of Science Policy: An Ontology-Based-Data-Management (OBDM) Approach

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

daraio@dis.uniroma1.it, lenzerini@dis.uniroma1.it, leporelli@dis.uniroma1.it, henk.moed@uniroma1.it
[1] DIAG, Sapienza University of Rome, via Ariosto, 25, Rome (Italy)

paolo.naggar@gmail.com,alessandro_bartolucci@fastwebnet.it
[2] Studiare Ltd.

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

The OBDM approach in a nutshell

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 is a form of information integration, where the conceptual model of the application domain, formulated as an ontology expressed in a logic-based language, replaces the usual global schema. The integrated view that the system provides to information consumers is not merely a data structure accommodating the various data at the sources, but becomes a semantically rich description of the relevant concepts in the domain of interest, as well as the relationships between such concepts.

Sapientia: a Platform for Developing Science of Science's Policy Models

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 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 (see examples in the following).

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. To represent the knowledge production activities, at an atomic level, we 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 can analyze 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 act as resources 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.

Conclusions

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. 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. 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.

Acknowledgments

Research support from the Progetto di Ateneo 2013 of the Sapienza University of Rome is gratefully acknowledged.

References

Daraio C., Lenzerini M., Leporelli C., Moed H.F., Naggar P., Bonaccorsi A. & Bartolucci A. (2015). Sapientia the Ontology of Multi-Dimensional Research Assessment*, Proc. ISSI.*

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

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

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