Ontology-mediated Data Integration and Access in Research and Innovation Policy

Alessandro Mosca,
SIRIS Lab, Research Division of SIRIS Academic, Barcelona, Spain
{*name.surname*}*@sirisacademic.com*

Keywords. Open Innovation ecosystem, OBDA/I, Data-driven policies

1. Research and Innovation policy making: a preamble

The goal of the European Research and Innovation (R&I) Policy is to help tackle the great challenges Europe is facing: spurring smart, sustainable and inclusive economic growth and job creation, building a resilient society while accommodating globalisation. Moreover, the current economic situation and the requirements of public accountability require the maximisation of the Union's budget's effectiveness, the capacity to demonstrate tangible results on the ground, and the relevance of funded research not only for scientific communities but also for the economy and society at large.

Unfortunately, the emerging disconnection between the fast changing and fast growing field of scientific research and the tools that the policy makers have at their disposal to measure and understand its current state, presents serious challenges. This threatens the effective potential to translate knowledge into socio-economic value (*Open Innovation 2.0*, Directorate-General for Research and Innovation, EU 2014). Decision-makers at universities, research institutions, companies, together with policy makers and the involved public administrations, all are part of a knowledge (learning, discovering and innovating) engine that, if well managed, could create wealth, jobs, growth and social progress (*The Knowledge Future*, Directorate-General for Research and Innovation, EU 2015).

The paper argues about the fact that the above scenario gives precise requirements to those who are involved in the design and development of digital technologies. The tools that the Commission, and the policy makers at different territorial levels, have at their disposal to measure and understand both the current state of the scientific research and the potentiality of the achieved results in fostering innovation, are still to be developed indeed, in order to:

- **define** of future research and innovation priorities (e.g. fields, technologies, sectors);
- **identify** of emerging fields and competitive solutions for transfer and investments
- **shape** of new policies by means of scientific and technological evidence;

- **re-adjust** of existing policies via evidence-grounded monitoring and learning processes.

2. Open innovation: a challenge for semantic technologies

Being able to implement innovative knowledge transfer channels and to improve the communication of scientific knowledge to a larger portion of the society is certainly one the most import message the European Commission (EC) is currently trying to spread into the scientific and entrepreneurial communities. However, the more we move in the direction of the current European policies and reccomendations, the more we realise that it is not only a matter of *accountability* and *knowledge transfer* in front of a larger audience: the relevant actors in the research and innovation audience are expected to assume new roles, duties and rights with respect to what we have seen happening during the last ten years.

On this respect, the concept of triple helix of university-industry-government relations emerged in the mid '90s [1], and has been further elaborated for explaining structural developments in knowledge-based economies [2]. This concept points towards the dynamics that can be expected as a result of interactions involving bi-lateral and tri-lateral relations among university, industry and government. The central issue of the triple helix is that the three helices, as well as the interactions among them, define the rules of the game of a place (e.g., a region or a nation), thereby constraining its development possibilities. Full recognition of these constraints is the first step to understand which development paths are viable and, eventually, how to politically intervene to modify them *in itinere*. Sustainable growth of the system requires that the helices actions are consistent. In fact, when the helices are out of alignment, imbalances occur.
More recently, a fourth helix - the collective sphere of civic societies and larger social networks - has been added to the three initial ones [3]. This latter definition makes explicit the role of non-university-industry-government organisations, such as civil associations or non-profit organisations and social enterprises, in shaping local and regional development paths. Quadruple helix processes and positive outcomes are rarely the result of undeliberate interactions. Collaborative and reflective schemes are needed, but they cannot be based just on records of past success of single stakeholders and traditional governance solutions, nor on optimistic declamations of nice common targets in public-private-partnerships. Learning by interacting, learning by monitoring and evaluating, and experimental solutions should be practiced deliberately, knowing that incredible power (and potential support) of the present digital technology. The concept of Open Innovation is indeed all about that.

The *Open Innovation, Open Science, Open to the World - A vision for Europe* document[1], recently published by the EC, represents the synthesis of the new European Union approach, based on quadruple helix collaborative design and work, and on the use of digital technologies, in order to:

1. **share information and knowledge beyond scientific publications**, and

[1] European Commission (2016). Open innovation, open science, open to the world- a vision for Europe. European Commission's Directorate-General for Research & Innovation (RTD), Bruxelles.

2. **support the co-design and co-planning of scientific policy and strategies that include the quadruple helix actors**.

As one would expect, the new vision brings new needs, new requirements, and a specific attention focused on two main elements: the *Public Engagement* - the users are in the spotlight: an invention becomes an innovation only if users become a part of the value creation process; and the *Ecosystem* - the creation of a well-functioning eco-system that allows co-creation becomes essential for Open Innovation. In such an eco-systems, the relevant stakeholders are collaborating along and across industry and sector-specific value chains to co-create solutions to socio-economic and business challenges.

Open innovation has therefore to be taken as the outcome of a complex co-creation process involving knowledge flows across businesses, academia, financial institutions, public authorities or citizens. Consciously intervening to tune these flows, while having in mind the challenges above, is owed to the taxpayer and to ourselves, and is tough. Responses will need to be based on theory and empirical evidence, as well as conveyed in a manner that must be understandable [4]. It is within this newly suggested conceptual framework that the digital technologies behind the design and the implementation of support platforms gain a specific characterisation: **they have to be open, capable of accommodating new strategic demands, new uses and new data sources, both internal and external**.

On December 7-8, 2007, thirty open government advocates gathered in Sebastopol (California) agreed on the following 8 principles characterising an Open Data Government initiative, currently adopted also by the EC. "Government data shall be considered open if it is made public in a way that complies with the principles below": *Complete*: All public data are made available; *Primary*: Data are as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms; *Timely*: Data are made available as quickly as necessary to preserve their value; *Accessible*: Data are available to the widest range of users for the widest range of purposes; *Machine processable*: Data are reasonably structured to allow automated processing; *Non-discriminatory*: Data are available to anyone, with no requirement of registration; *Non-proprietary*: Data are available in a format over which no entity has exclusive control; *License-free*: Data are not subject to any copyright, patent, trademark or trade secret regulation.

The reference technologies in this context have been clearly stated. They have to follow the EC "Linked Open Data" standard, clearly introduced in the EU policy "A Digital Agenda for Europe"[2]. In the document, Linked Open Data are introduced as the current standards to represent data on a wide range of topics which makes it easier for developers to connect information from different sources, resulting in new and innovative applications: Linked Open Data enables, as said there, a "browsing" or "discovery" approach to finding information, as compared to the usual "search" practice[3]. The formal languages behind the concrete realisation of a Linked Open Data initiative are the well known standards: RDF ("Resource Description Framework"): the flexible data model based upon the idea of making statements about resources in the form of subjectpredicateobject expressions, known as triples; RDFS/OWL2 ("Resource Description Framework Schema"/"Web Ontology Language"): the schema and ontology languages for describing concepts and relationships; SPARQL (SPARQL Protocol and RDF Query Lan-

[2]https://europa.eu/european-union/file/1497/download_en?token=KzfSz-CR
[3]https://data.europa.eu/euodp/en/linked-data

guage): the query language RIF (”Rule Interchange Format”): a rules language originally designed to exchange rules among different existing rules dialects; RDFa (”Resource Description Framework in Attributes”): the language for marking up data inside HTML-based Web pages); and HTTP communication protocol (”Hypertext Transfer Protocol”): the application protocol for distributed, collaborative, and hypermedia information systems, at the foundations of the so-called World Wide Web.

2.1. Ontology-based data management

The above mentioned standards are usually referred as *Semantic Web technologies*. Semantic Web technologies are formal languages and solutions that bring structure and meaning to information, that adhere to the specific set of W3C open technology standards. The languages and technologies introduced here are the languages that the open innovation and open science platform must rely on for the design and implementation of their data integration and data access services, according to the present EC recommendations, guidelines and visions: the accomplishment of these requirements would ensure the platforms to be compliant with the directives of the EU about Open Innovation and Open Science.

The management of complex kinds of information has traditionally been the concern of Knowledge Representation and Reasoning (KR&R) in Artificial Intelligence. In particular, a recently introduced paradigm that combines the possibility of using reasoning with respect to domain knowledge encoded in an ontology, with a mechanism to use the same ontology also for high level, integrated access to data sources, is that of Ontology-Based Data Access and Integration (OBDA/I) [5]. Ontologies are usually specified in Description Logics (DLs) [6], a family of knowledge representation languages that provide one of the main underpinnings for the OWL Web Ontology Language as standardised by the W3C[4]. DLs are equipped with a formal semantics based on First-Order Logic. This formal semantics allows humans and computer systems to exchange DL ontologies without ambiguity as to their meaning, and also makes it possible to use logical deduction to infer additional information from the facts stated explicitly in an ontology. In the OBDA/I setting, the most commonly used language is OWL 2 QL[5], which is the profile (i.e., sub-language, in W3C terminology) of OWL 2 that is specifically tailored for efficiently querying large amounts of data. The domain ontology is then connected to the data sources through a declarative specification given in terms of mappings that relate symbols in the ontology (*concepts* and *properties*) to views over the data expressed by means of SQL queries. The ontology and mappings together expose the data in the sources in the form of an RDF graph, which however is not materialised. Queries, which can be formulated over the concepts and properties of the ontology, are interpreted over a virtual RDF graph, and are translated, making use of the mappings, into SQL queries over the data sources.

In an OBDA/I setting, users simply query the ontology, and no longer need an understanding of the data sources, the relation between them, or the encoding of the data. Due to the presence of an ontology, and of explicitly defined mappings, OBDA technology facilitates the access and the SPARQL-based exploration of the integrated data, especially when non-technical end-users are involved. Fig. 1 shows a high-level representation of

[4] `https://www.w3.org/OWL/`
[5] `http://www.w3.org/TR/owl-profiles/`

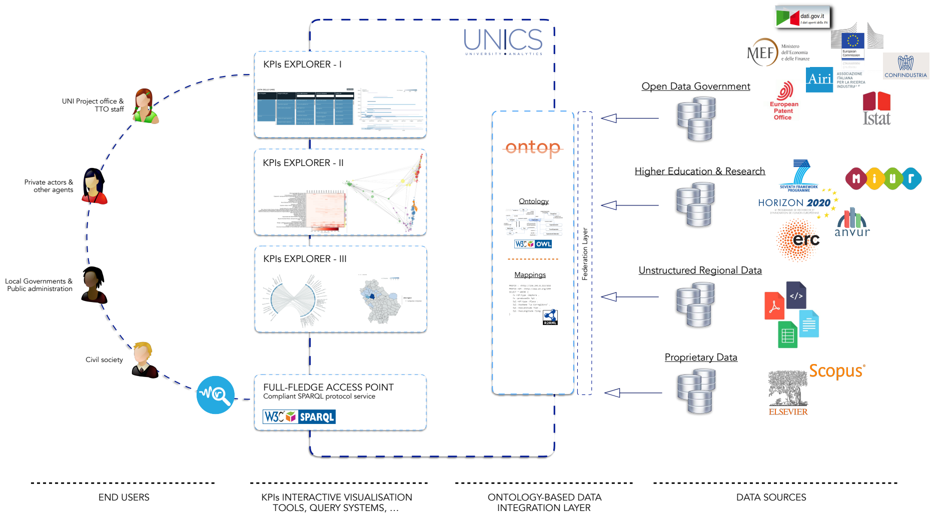

Figure 1. UNiCS (`http://university-analytics.com/`) platform architecture tailored for the Tuscany's Observatory of Research and Innovation. In the right-hand side of the image, the list of repositories to be integrated, currently organised according to the Academia, Technology & Innovation, Health care, Public Policy Making, and Social Sciences and Humanities scenes. The central part of the figure points to the two major components of the platform: the master ontology and the source(s)-to-target mappings. On the left hand side, the front-end of the platform, made out of data visualisation tools which are designed and implemented to answer the information needs of distinct users, and the SPARQL endpoint.

an OBDA/I system for data integration and access developed by SIRIS Academic in the context the Tuscany Region "Regional Research and Innovation Observatory" project in Italy, is presented.

3. Concluding remarks

The introduced architecture represents only one possible exploitation of semantic web technologies and principles to support the current EC vision and strategy on the Research and Innovation policy making. OBDA/I technologies support the actors of the quadruple helix, who are usually neither computer scientist nor database experts, in looking for interesting correlation and/or patterns in the data, especially when the data are coming from a multitude of disparate, originally not-homogeneous, data sources. More concretely, OBDA/I can be used by private and public R&I actors to get:

1. an exact and detailed **map of their own current state**, including internal processes, human resources, skills and research portfolios, technological and economic strengths and weaknesses;
2. extensive **knowledge of the context** in which they are operating, including needs and requirements of their stakeholders and competitors strategic profiles;
3. a robust **decision-making process** that ensures priorities are informed and recognised internally as legitimate.

It will be, of course, in charge of the policy makers to then translate the insights coming from the data into applicable and reasonable political actions (such as, research and innovation investments). Here, we simply tried to convey the message that the introduced vision and strategy about R&I policy making, may represent an opportunity to further support the research activity in KR&R, and the consequent development of OBDA/I and semantic technologies, in the next few years. Strictly speaking, rather than assuming a passive role and spending further energies in devising '*Cahier de Doléances* on the actual budgetary austerity' in the academia, we strongly suggest to opportunistically point out the pivotal role the semantic- and ontology-based technologies can play in such an arena, where a multitude of information sources and relevant datasets have to be identified, consistently integrated, and accessed, for instance, by policy makers, stakeholders, and domain experts who are not computer scientists. Over the last few years, a multitude of collaboration raised with the objective of developing ontology-based platforms for data integration and access in the Italian, Spanish, and French policy making arenas (see, for instance, [7] and [8]), at different organisational levels. All of them represent successful experience in the application of semantic technologies for supporting policy making in research and innovation. Nonetheless, chances to get the 'ontology-based' message heard by the European Commission itself are real nowadays, and the playground is open by default to all of you, experts in the knowledge representation and semantic technologies field.

References

[1] H. Etzkowitz and L. Leydesdorff: *Universities and the global knowledge economy: a triple helix of university-industry-government relations*. Amsterdam: University of Amsterdam, 1995.

[2] H. Etzkowitz and L. Leydesdorff: *The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of universityindustrygovernment relations*. Research policy 29(2): 109-123, 2000.

[3] E.G. Carayannis and D.F. Campbell *'Mode 3' and 'Quadruple Helix': toward a 21st century fractal innovation ecosystem*. International journal of technology management, 46(3-4), 201-234, 2009.

[4] J. Lane, *Assessing the Impact of Science Funding*, Science 2009.

[5] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, M. Rodriguez-Muro and R. Rosati, R, *Ontologies and Databases: The DL-Lite Approach*. Reasoning Web, 5689, 255-356, 2009.

[6] F. Baader (Ed.), *The description logic handbook: Theory, implementation and applications*. Cambridge university press, 2003.

[7] N. Antonioli, F. Castanò, C. Civili, S. Coletta, S. Grossi, D. Lembo, M. Lenzerini, A. Poggi, D.F. Savo, E. Virardi, *Ontology-Based Data Access: The Experience at the Italian Department of Treasury*. CAiSE Industrial Track 2013: 9-16.

[8] A. Mosca, B. Rondelli, G. Rull, *The OBDA-based "Observatory of Research and Innovation" of the Tuscany Region*. In Proceedings of The Joint Ontology Workshops, JOWO 2017 (WS-CEUR proceedings). To appear.