Towards an Open Quadruple Helix Indicators Factory

Andrea Bonaccorsi[1], Giuseppe Catalano[2], Cinzia Daraio[2], Henk F. Moed[2]

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

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

This version: 25 July 2016

Table of content

Abstract

This paper proposes a new approach to data integration in R&I that radically departs from the traditional silos-based approach. It illustrates the potential of this approach for the realization of a new Quadruple Helix interactive innovation model and for the harmonization of economic and financial accounting data of European universities. The expected impacts, in terms of sound evidence-based policy making are also discussed in the concluding section.

1. Introduction

In science, technology and innovation policies there is a strong need to improve the quality and intensity of use of data in order to make better decisions. The push towards *analytic government*, advocated by President Obama, goes in this direction. In the same direction the OECD (OECD 2015a; 2015b), the European Parliament (EPRS, 2014) and the European Commission are working around the notion of *open science*.

At the same time it seems clear that the goal of accessing and using big data in ST&I policies may be hampered by the adoption of an obsolete approach to the architecture of production and collection of data. In essence, in ST&I we witness:

- Separate databases regarding different aspects of science, technology and innovation (e.g. publications, patents, funding, webometrics, entrepreneurship, regional growth);
- Unsolved issues of disambiguation;
- Lack of an official census of actors (universities, public research organizations (PROs), companies).
- Poor geo-referentiation of published data.
- Lack of breakdown of data, for instance, by territory, scientific discipline or technology.

Policy makers, on the other hand, perceive increasing needs for data that are *granular* (by territory, scientific field, technology), *cross-referenced* (for example, combining publications and patents for all universities), and *interactive* (allowing feedback between data users and producers)[1].

Faced with these problems the approach commonly adopted by policy makers is to start new initiatives aimed at integrating datasets, by creating new databases and using a "silos" approach to data integration. For example, if a government aims to understand the impact of public research on innovation using patent data, it must demand (and fund) a custom-made study that integrates data sources on the publication and patent silos, solving *ex novo* a large number of issues of disambiguation, errors, compatibility and comparability of data. New databases created with the silos approach are expensive, yet they are not interactive, inter-operable, updatable, and scalable. They become rapidly obsolete. A new cycle of data integration will be, sooner or later, needed again.

We suggest a radical departure from this approach, sustained by state-of-the-art technology. We have already achieved a good level of maturity along this direction, although much work is still to be done. We report on the results achieved thus far and on a timescale to achieve the final goal.

2. Proposed approach

The radical departure we suggest is based on two pillars: a new approach to data integration and management, and a systematic effort to build so called Authority Files.

With respect to data, we advocate the adoption of an Ontology-Based Data Management (OBDM) approach (Poggi et al., 2008; Calvanese et al. 2009). This technology is state-of-the-art, has several applications in the business and government sectors, and has been only very recently proposed in

[1] See Daraio and Bonaccorsi (2016) for a discussion.

the field of ST&I (Daraio, et al. 2016a,b). An OBDM system is based on a three-level architecture, constituted by[2]:

- The *ontology*: a conceptual, formal description of the domain of interest (expressed in terms of relevant concepts, attributes of concepts, relationships between concepts, and logical assertions characterizing the domain knowledge, supported by a formal language).

- The *sources*: 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*: a precise specification of the correspondence between the data contained in the data sources and the elements of the ontology.

In Daraio et al. (2016a) we have introduced the OBDM approach in order to coordinate, integrate and maintain the data needed for ST&I policies. This paper summarizes the main advantages of OBDM with respect to the traditional silos-based approach to data integration, namely: *conceptual access to the data*, *re-usability*, *documentation and standardization*, *flexibility*, *extensibility and opening of the system*. Daraio et al. (2016b) focus on three main advantages of OBDM in data integration for research and innovation analysis, namely *openness*, *interoperability* and *data quality.*

An interdisciplinary project, funded by Sapienza university of Rome, has developed *Sapientia, the ontology of multidimensional research assessment,* which is formed by more than 500 concepts organized in 10 modules. It is currently being completed. Its consolidation and testing is underway. At the same time, it benefits from the development of a logic-based reasoner, called *Mastro*, which allows the automatic creation of queries to be executed at the level of data, on the basis of queries formulated at the higher level of concepts in the ontology-layer (Calvanese et al. 2011). This is an extremely powerful technology allowing the construction of new indicators, without the need to build up integrated datasets. In other words, the system may query whatever heterogeneous and independent source of data, on the basis of the formulation of user needs based on concepts.

In this approach, if a government is interested in integrating data on publications and patents, the system will run an abstract definition of, say, “country”, “university”, “article”, “patent” and the like, and will automatically recognize their linkages. At this point the issue of linking all articles and all patents to the same entity (university) is formally specified. Then the *Mastro Studio* system transforms this problem into a formal query to be executed on separate databases. This operation is possible because there is a unique definition of “university” at the ontology level, and there is a complete list of all the possible occurrences of the names of any university in the given country in all existing databases. No full scale disambiguation will be needed. This is made possible by our second challenge to the established approach- Authority Files.

[2] The presentation of the OBDM approach follows Daraio et al. (2016a,b). For a more technical presentation, see Daraio et al. (2016c).

In fact, the second pillar of the approach is the construction of *new Authority Files*. In database language, an Authority File is a structured list, associated with formal procedures for its maintenance and updating. Our approach requires a maximum emphasis on the construction of officially validated lists of entities of the ST&I systems all over the world.

In the current phase, three main types of Authority Files are considered for inclusion in *Sapientia*:

- Higher Education Institutions (HEIs)
- Public Research Organisations (PROs)
- Private institutions / companies.

All existing standards already in place worldwide (e.g. geographic references, ZIP codes etc.) must be associated to these Authority Files. At a later stage, other types of authority files could be added.

This choice is encouraged by a number of feasibility tests already carried out with success. In Europe, an official list of HEIs has been established in collaboration with all National Statistical Authorities, following the pioneering activities of the European projects Aquameth and Eumida, under the current project ETER, a project funded by the European Commission. Official lists are also available for US, Canada, and most Eastern Asian countries. UNESCO sources can also be added to the picture.

With respect to PROs the situation is more complex, because there is a need to ask individual actors to validate the list and the geo-referentiation of information at sub-units level (e.g. location of institutes or laboratories in multi-site large PROs).

With respect to companies, at least two matching schemes between names of companies appearing in patents and names of companies appearing in commercial sources are available.

All entities in the Authority Files are geo-referentiated at regional and possibly urban level. Once created the platform, further specification of Authority Files can be proposed. Other Authority Files might follow the institution level (e.g. funding agencies, governments, ONGs) or the individual level (e.g. authors of publications, inventors of patents). What is crucial at this stage is that Authority Files are designed, on the basis of the OBDM approach, in such a way to be compatible and inter-operable with other lists, that are candidate for future integration.

In our approach the Authority File must be:

- managed by a public authority (or delegated to an institution working on behalf of a public authority);
- updated regularly, authorizing the entry of new members (e.g., new universities);
- publicly available and subject to scrutiny and comment;
- associated to a publicly available list of all occurrences of names in any available source that can be disambiguated *in advance*.

The Authority Files ensure the smooth functioning of the OBDM. The availability of a validated and updated list makes it possible to integrate all kinds of data on these entities, facilitating the automatic matching of variant names in data sources with one of the occurrences associated to the list. This makes it possible to build up new indicators in an automatic, or semi-automatic way.

A distinctive feature of this approach is that the ontology of entities included in the Authority Files already include all logical aspects of entities that might be integrated in the future. An important example is individuals: while we still do not have officially available lists of, say, active scientists, or patent inventors, it is mandatory that the ontology already includes all logical aspects of the entity "scientist" or "author", or "inventor", so that in the future any advancement in the availability of data may be immediately exploited.

It has to be underlined that:

- the current feasibility studies very much focus on EU institutions, but the developed framework allows the coverage of non-EU institutions as well;
- the development of authority files for subject classifications and their concordance, enabling one to perform analyses by subject field combining data from distinct types of data sources (e.g.,publications, patents, funding) is a relevant task to be performed;
- the approach allows for standardization of concepts, data sources and analytic tools in quantitative science and technology studies, the need of which is often stressed both by experts in the field and by users of bibliometric/ informetric data (Daraio and Glanzel, 2016).

3. The power of Authority Files for scalability and inter-operability

A crucial methodological choice of our approach is the adoption of the concept of Authority File. It is important to explain why this is the case. In the quest for the benefits of information society, there has been too often the assumption that Big Data can do the job of providing information that policy makers can directly use. This is not always the case.

What we need is a layer of entities whose ontology and existence is relatively stable over time, so that its demography, whatever the turbulence it may have, can be traced. In abstract terms, this requires to identify a layer in which the rules for the inclusion of entities are unambiguously spelled out, the events of entry, survival, exit are highly visible, and the events have a timescale which is compatible with the timescale of observation. We suggest that this layer cannot be, for the time being, that of individuals, but must be that of organizations. This is currently possible for HEIs and companies, and will be possible, with a little extra investment, for PROs. These three entities will be the backbone of the OBDM.

In addition, we need a layer that would permit, in the future, the inter-operability with other Authority Files. We are deeply persuaded that in the near future the coverage of individual-level

information systems will increase to such a point that they might be used as a basis for the construction of indicators. Individual-level information on publications and patents are already available in a number of sources. Recent papers by Julia Lane and co-authors show the enormous analytical power of integrating individual-level information on higher education with data on earnings. At the same time, without a clear linkage between individual-level data and the organizational layer, there is no hope for the former to become a standard for making decisions. If we may download free information on the publications of N researchers at the University of Helsinki, but we do not know anything about the academic staff of the University of Helsinki, their number, disciplinary composition, and dynamics over time, and we do not have any comparison with other organizations, how can the individual-level information be used?
We suggest that the best thing to be done for the time being is to design the organizational layer (HEIs, PROs, and companies) in such a way to be already inter-operable with future developments in the individual-level sources of data. The organizational layer will catalyse the growth of individual-level sources. By incorporating all existing standards for data at individual level, the OBDM system will lay the ground for future growth.

Our choice of adopting the organization level does not come out of the dark. It is based on previous research and on the recognition of existing successful initiatives. In the last few years, several initiatives at European level have been based on an intense production and use of new data.
In the field of data on Higher Education Institutions (HEIs), the pioneering efforts of Aquameth (Daraio et al., 2011; Bonaccorsi and Daraio, 2007) and subsequently of Eumida (Bonaccorsi, 2014) have been transformed in an institutional initiative called ETER (European Tertiary Education Register), which has made publicly available microdata on universities in 2015 and 2016 and will publish new data in 2017.
In the same field, the mapping of diversity of European institutions (Huisman, Meek and 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 and 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). Two large ERA surveys were run in 2013 and 2014 and the corresponding datasets were published online (http://ec.europa.eu/research/era/eraprogress_en.htm).
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 are based on the construction of "partial" ontologies[3].

We suggest that the OBDM approach, coupled with a careful incorporation of all relevant partial ontologies and standards, can open the way to a new era in the construction of indicators.

Let us see the way in which the new approach may help to address some emerging challenges.

4. Is the current "indicator factory" fitting for a new Quadruple helix interactive innovation model?

The exponential increase in the availability of data and the impressive developments of tools for data mining and intelligence create huge opportunities to deliver the promise of the information society. According to some authors, these trends will lead to a fundamental change in the innovation model of advanced societies. This change has been described in a variety of ways, from the Open innovation paradigm (Chesbrough, 2006), to the Quadruple helix model (Carayannis and Campbell, 2009; Leydesdorff, 2012).

These authors suggest that there are two new dimensions to be added to the innovation model: (a) inclusion of citizens and their organizations, in a user-centered innovation model; (b) emphasis on a practice-based, interactive, bottom-up learning model, making large use of cross fertilization of ideas and leading to experimentation and prototyping in real world setting.

These conceptualizations try to make sense of the experiences made in the last 10-15 years in the way in which big Societal Challenges have been addressed in many advanced societies. Issues such as sustainability, climate change, urban congestion, mobility, or patterns of energy consumption have been addressed with a mix between research, Information Technology, and participatory approaches to social change. Citizens are much more and better informed than before, due to the data revolution, and share their experiences in digital communities. They increasingly ask to be involved in decisions. Innovation becomes a joint product between research, digitalisation, and social creativity.

It is important to recognize that this trend creates a challenge to the old model of ST&I indicator construction and use. The three main actors of the Triple helix (government, academia and industry) shared a similar approach to indicators and to their use in decision making. While they

[3] Consider, for instance, the following:

- ORCID (http://orcid.org/).
- CERIF (http://www.eurocris.org);
- CASRAI (www.casrai.org);
- ISNI (www.isni.org),
- Ringgold (www.ringgold.com)
- CODATA http://www.codata.org
- VIVO (http://www.vivoweb.org/).

differ significantly on a large number of dimensions, still they know how to produce empirical evidence, they appreciate (perhaps partially) their statistical significance and use indicators as a justification for their decisions. The overall OECD ST&I indicators framework was perfectly suited for these actors. Accurately defined, regularly collected, internationally comparable, and statistically validated indicators have been (and still are) an essential element of the decision making process. Indicators follow a clear-cut distinction between domains of activities.

The fourth actor in the helix follows a somewhat different logic. Citizens mobilize around specific issues. These issues often cut across traditional boundaries: they call for multidisciplinary knowledge, involve public-private interaction, need radically new business models and/or public governance models. The established factory that produces ST&I indicators is not adequate here. The required indicators are often new, must be created ex novo in order to illuminate complex issues. They cut across existing domains of indicator production. They must be designed and produced interactively.

In the current scenario, each time a policy problem requires the integration of data from multiple sources there is a distinctive need for a dedicated effort, usually from a research team or consultancy unit. Our approach will allow policy makers to build up their own indicators and obtain results that integrate in a reliable way heterogeneous data. It will permit policy makers to build up indicators on-demand. This will be critical for addressing issues that go beyond existing classifications, such as Key Enabling Technologies (KET) and societal challenges (SC). New definitions can be entered the system and can be used to address the platform, construct new indicators, and produce original information.

How the proposed OBDM approach can address this challenge?

We propose a radically innovative conceptual model for the generation of indicators, which can be described in five steps (this approach follows and extends to a wider context the approach proposed in Daraio and Bonaccorsi (2016) on which the following presentation is based on).

The first step is conceptualizing education, research and innovation systems as complex systems formed by more elementary entities. There are several possible ways of doing this, as proposed in the literature. One distinction is between an actor-centred perspective, which defines elementary action units in the system, and a function-based perspective, which, in contrast, works more on structural relations and flows within the system. We will try to reconcile these two perspectives. A starting point for conceptualizing this issue is to think in terms of actors[4].

*The second step is to develop concepts associated with entities***.** The development of concepts will be enriched by the consideration of users' requirements for indicators. Rather than following strict definitions, actors should be conceptualized in terms of a list of associated concepts. For example, a university is associated with concepts, such as students, professors, disciplines, location, publications, etc. Each of these concepts is, in turn, susceptible to further conceptualization. For example, students are associated with concepts, such as nationality, country of their last degree,

[4] See Daraio (2015).

gender, field of education, etc. For each concept an extensive assessment of the underlying definitions and of the relations with the other concepts will be carried out. It is not important that definitions are formally similar, since the higher level concepts will be generated by variable degrees of membership of the lower level concepts. This is a crucial point. What is needed in the traditional integration of databases is a perfect matching between the concept and the definition in order to carry out a query.

The third step is to look for lists and standards, or, more generally, in database language, Authority Files. The existence of a list is enormously important, since it allows exhaustive searches. Lists are flexible tools, in the sense that they are associated with procedures for updating and correcting. For example, the availability of an Authority File of companies investing in R&D, or a list of higher education institutions (HEIs) or of public research organizations (PROs), is an important step. Standards are also extremely important. For example, one would take on board standard definitions adopted at an international level as to what constitutes an author of scientific publication (ORCID, see www.orcid.org), or a research organisation (CERIF, see www.eurocris.org).

The fourth step is to establish links between concepts. Once these steps are carried out, what we obtain is a graph of actors, each with an associated graph of concepts (direct and indirect link). The links between actors are represented as links between concepts associated to actors. Then we are ready to implement a principle for the integration of education, research and innovation information systems. The informal principle we follow is: *put your information in the region of the graph in which the potential for generation of other information is greater*.

In many cases there will be a clear need for delineation or profiling work, in particular in the identification of new scientific areas or technologies. This is not an automatic task. However, the OBDM approach allows the inclusion of most advanced technologies of machine learning. Correspondence tables between concepts can be integrated and exploited (e.g. correspondence between subject categories of publications, patent classes, industrial classifications). Thesauri of words from publicly available studies in bibliometrics and scientometrics can also be integrated.

The final fifth step is the mapping of the datasets to the ontology.

Suppose there is a public debate on a new technology, followed by controversies on their desirability or feasibility (say, the civil use of drones, or the potential of new batteries for electric cars).

After the construction of the information architecture described above the actors, following a preliminary (but not expensive thanks to the pre-processing available in the architecture) profiling of the field, might:

- extract metadata on all universities, PROs and companies publishing papers in the field, and/or producing patents;
- examine the distribution of these actors (by world region, country, size, type) and the distribution of their production;

- apply tools in data intelligence aimed at extracting trends and identifying promising research and innovation directions;
- build up indicators of concentration, entry/exit, size, in order to understand the life cycle of the technology;
- compare the technology with others, following an analogical modelling;
- examine the patterns of funding of research (perhaps only partially);
- build up a list of active players in the field (HEIs, PROs, companies) to be traced over time in order to detect changes in the composition and activities.

All these elements, and possibly others, might contribute to the construction of new indicators. After experimentation and learning, some of them might become candidates for a regular construction and publication.

In the near future, these might be linked with individual-level sources of data.

5. An example of usefulness of this approach: Harmonizing economic and financial accounting data of European universities

In this section we report an example of potential application of the proposed approach, to develop comparable and harmonized performance indicators across international higher education institutions.

In the current situation of decreasing availability of public budgets for universities, the financial sustainability of universities goes hand in hand with their accountability, and the diversification of their income sources (Esternmann, 2011, 2013). The relevance of this topic is also significant from a *cost-sharing* perspective, that is in view of the increasing institutional autonomy of universities who diversify their sources of funding both as regards tuition fees and the supply of goods and services (third mission activities).

Within this context, accountability and efficient and effective management of assets and resources appear as crucial aspects for maintaining the long-term sustainability of universities' activities. Indeed, the introduction of accrual accounting in the public sector has been one of the main business-like devices of New Public Management (Agasisti, Arnaboldi and Catalano, 2008).

Moved by these considerations, several governments in Europe introduced reforms to their accounting rules to offer university managers the opportunity of strategically manage their resources. As an example, Agasisti and Catalano (2013) describe how the Italian legislation followed international experiences in order to compare universities' balance sheets and economic/ financial outcomes. The Italian government Decree No. 18/2012 requires universities to adopt accrual accounting. For a technical presentation, see Agasisti et al. (2015) who, confirming the results of previous literature, find a low level of compliance to full accrual accounting principles. Their analysis reveals the scarce contribution of International Public Sector Accounting Standards (IPSASs) to promote full accrual accounting in the public sector. Indeed, while there are

not specific accrual accounting standards for the higher education system, some international organisations (such as the OECD or the International Federation of Accountants) claim for the use of IPSASs as main guidance tool to ensure the effective use of accrual accounting in public sector. Another risk of this reform relies on the ability of administrative staff to manage the changes. Moreover, a system of good internal relationships needs to be created to guarantee that departments maintain their managerial autonomy in the context of a unitary accounting system (Agasisti and Catalano, 2013).

More recently, Brusca, Caperchione, Cohen and Rossi (2015) compared public sector budgeting, accounting and auditing systems in 14 European countries. Their analysis shows that budgeting and accounting systems have a "significant heterogeneity between countries for all government levels and that there is also a lack of harmonization among different government levels within countries. In most countries, accounting standards are different for central, regional and local governments. Furthermore, although in all the countries analysed there are provisions for both internal and external audits, auditing in the public sector displays a heterogeneous panorama." They conclude with a view of the readiness for change to IPSAS or EPSAS in the countries analysed, showing important differences in the challenges and efforts necessary to move in that direction.

However, from a state of the art investigation of the harmonization of public sector accounting (see Caperchione, 2015) public administrations are facing two important and interrelated challenges, that are the reduction of the distance between their accounting systems and the choice of an appropriate set of standards to accomplish this goal.

In this respect, Caperchione (2015) identifies many open issues: "the willingness to adopt international standards is not spread equally, and is sometimes quite limited; there is more than one set of standards available; and a general consensus on which set would best fit the needs of EU member states is still missing. The debate will probably last for a long time, and will allow for the refinement of the proposed solutions; but none of them, we believe, will prove satisfactory, if a wholly uniform system is designed which overlooks the importance of good management at a decentralized level (Caperchione, 2015, p. 9)".

There is an increasing evidence that the adoption of digital technologies by organizations not only affects the economics of operational and managerial processes but also mobilizes extensive social and organizational effects. Digitization impacts the form, substance, and provenance of internal accounting information with attendant consequences on the behaviour and actions of organizational participants and on the functioning of enterprises more widely. Knowledge about the influence of the deployment of digital technologies on management accounting thinking, processes, and practices is increasingly available. Digitalization and accounting changes where already analysed, for instance, in Bhimani (2003).

Our proposal is to experiment an Ontology Based Data Management Approach, using Sapientia as a base, for the harmonization of heterogeneous accounting systems at European/international level, along the lines of, *mutatis mutandis*, what has been done in the financial accounting

reporting (see e.g. Spies, 2010, Wenger et al. 2013). This would be an interesting new promising research line to pursue in the next decade.

The comparison of universities' budget information, as the Italian experience shows, requires an accounting system of economic and financial kind, since the universities are producing goods and services, even when they are sold in the market, with significant public contribution, and make investments in real estate and research equipment. Uniform and dedicated reporting formats, but especially shared accounting principles, are needed.

This example shows the potential of our approach with respect to the existing ones. While in a traditional database approach there would be a huge work of comparison of data, in the OBDM approach the most important aspect is to develop a conceptual ontology that captures the main abstract elements of indicators. Then the use of data by policy makers and users will be associated to clearly defined and conceptually organized data without having to consider the data technicalities and limitations.

To make an example, once the ETER dataset has been created, it has been possible to start a number of pilot projects to build up comparable indicators on universities funding and expenditures in Europe. It turns out that this task cannot be achieved moving upward from the definitions given in various countries by administrative authorities and with the limited resources available within the project. What should be done, in our perspective, is to develop a higher level ontology to characterize the main features of the funding and expenditure activities. Based on this ontology, the data that fully meet the definitions at the ontology layer can immediately be used; those that have a partial overlap with the definitions can be used with qualifications.

We think that *Sapientia* and OBDM are promising technologies to be exploited for defining economic and accounting principles that can be the prerequisite of collecting comparable data even if arising from accounting systems and procedures with different characteristics. We believe that the work done and the experiences realized with *Sapientia* and the technology of OBDM represent a promising avenue that need to be explored to develop further research to address the comparability problem of economic and accounting data of universities.

6. Expected impacts: towards a sound evidence-based policy making

There are several dimensions of impact of our proposed approach to STI indicators building with respect to the issue of evidence-based policy making.

The first is the *integration of data* to an unprecedented level. Given the advanced methodology proposed, it will be possible to integrate heterogeneous data while preserving their quality and integrity, so that they can be reliably used by policy makers and stand credibly in the face of Parliaments, media, and public opinion. These data can be integrated with information on

economic activities, sociocultural initiatives, infrastructures, and any other activity at regional or, in some cases, local level. Thus, *data integration* is the first contribution to evidence-based policy making.

The second dimension of impact on policy making can be defined in terms of *user-friendliness and accessibility*, in the direction of a Quadruple helix innovation model.

The third dimension of impact is in terms of *scalability*. Once constructed, the platform will allow the expansion and integration of new data sources. If these data are compatible with the ontological definitions, they can be immediately integrated. If they are not compatible, for example because they introduce completely new entities (e.g. new types of actors), then a formal procedure at the level of ontology will provide full definitions and will automatically generate all profiles of compatibility (or lack thereof) with the current data. This is an enormous gain in terms of usability and benefits for policy making.

A crucial assumption of our approach is that we are moving into a world in which data are not only largely accessible, but also self-produced by actors of the social system. This is not a neutral statement, but one that is definitely more likely to be realized in the world of science, and to a certain extent, of technology.

The system we call for is open, its underlying software will be Open Source, together with all metadata, publicly available data will be linked and structured. We anticipate that this will increase the use and capitalization of existing information. The implementation in the platform of the most advanced computational techniques (for disambiguation, geo-referentiation, location, crawling of web sources, data quality, data integrity and the like) will give it a self-propelling thrust, in the sense that actors themselves will sustain the production and validation of data.

Table 1 in Appendix summarizes existing research and innovation activities that could be coordinated within such a broad indicator factory design approach we propose.

Once established a robust platform for institutions, it will be possible to assign their inputs and outputs to geographic units, following transparent procedures (associated to clearly defined estimates of measurement errors). We focus here, for instance, on regional studies. It will be possible to link STI data to regional data to an unprecedented level of integration, including:

- Regional covariates coming from Eurostat rich regional data (see the available data at http://ec.europa.eu/eurostat/publications/regional-yearbook)
- Data on migration established by OECD (http://www.oecd.org/migration/mig/dioc.htm) (see also several sources at http://perso.uclouvain.be/frederic.docquier/oxlight.htm), with particular reference to migration of skilled workers and graduates
- Data on territorial infrastructure and activities from ESPON (http://www.espon.eu/main/)

By allowing the transparent integration of these datasets, countless opportunities will emerge to explore issues of regional interest.

We claim that an intelligent integration of existing data in an open-linked data platform may permit the construction of new indicators, without having to design the indicators on a custom basis. Following our proposal, no data should be collected without a clear definition of the logical structure, but once the data are collected this way, the construction of new indicators is greatly facilitated. Table 2 in Appendix summarizes the ways in which the policy needs of several actors in the ST&I systems can be addressed with the new approach.

The policy impact is manifold and durable. There is a need for investing into an infrastructure, whose benefits will be reaped for decades.

Appendix

Table 1. Examples of information that can be integrated in our proposed architecture

Related activity	Description	References and/or websites
A. Research and innovation activities at the level of institutions		
A.1 Higher education institutions		
European Tertiary Education Register (ETER)	Register of all European Higher Education Institutions (HEI)	http://eter.joanneum.at/imdas-eter/suche/erweitert-liste.jsf.
U-Map	University multidimensional mapping	www.u-map.eu
OECD	Knowledge Triangle Project	Internal documentation
National Science Foundation	Institution Rankings + Higher Education Research and Development (HERD) Survey	http://ncsesdata.nsf.gov
US Government	Integrated Postsecondary Education Data System (IPSED)	http://www.data.gov/education/
Center for Measuring University Performance	Data on funding	http://mup.asu.edu/research_data.html
Patent data	Dataset on academic patenting	http://www.esf-ape-inv.eu
RISIS	Research infrastructure for research and innovation policy studies	www.risis.eu
A.2 Public Research Organisations (PROs)		
Eurocris	European initiative aimed at standardizing data on research funding	http://www.eurocris.org/
A.3 Companies		
Corporate patent data	Various matching schemes	http://epo.org
Corporate R&D data	Corporate R&D Scoreboard CONCORDi	at http://iri.jrc.ec.europa.eu/scoreboard15.html http://iri.jrc.ec.europa.eu/concord/2015/index.html
B. Research and innovation activities at the level of individuals		
B.1 Disambiguation of inventor names	Research developments in the field of matching schemes and disambiguation of inventor names	http://cahiersdugretha.u-bordeaux4.fr/2012/2012-29.pdf). http://funginstitute.berkeley.edu/wp-content/uploads/2015/08/AutomatedDisambiguation-of-US-Patent-Grants-and-Applications.pdf http://arxiv.org/pdf/1601.01963v1.pdf
B.2 Mobility of inventors	Mobility identification and tracking	http://www.wipo.int/econ_stat/en/economics/publications.html
B.3 U-Metrics	Integration of administrative data	http://www.cic.net/projects/umetrics
B.4 Disambiguation of authors	Author identification	http://orcid.org
A. **Research and innovation activities at thematic level**		
C.1 Geographic level: country and region	Integration with regional covariates	http://ec.europa.eu/eurostat/publications/regional-yearbook http://espon.eu/main
C.2 Patterns of specialization and co-specialization	Revealed Technology Advantage (RTA)	https://ec.europa.eu/research/innovation-union/pdf/technological-specialization-of-countries.pdf
C.3 Thematic level: new and emerging	Definition of query schemes and retrieval of georeferentiated data	

research fields and technologies		
C.4 Network studies	Definition of flow data and mapping	

Table 2. Examples of policy making issues that can be addressed with our proposed architecture

Actors	State of the art and unsolved issues	Proposed solution by our approach
National policy-makers	Custom-made indicators require ad hoc studies and are expensive Indicators are not granular (e.g. the breakdown by region or by field is not available) Indicators cannot be cross-referenced (e.g. publications and patents) Data on scientific publications do not include PROs appropriately Publications data in WoS differ substantially from data in Scopus and the interpretation is difficult	Construction of custom-made indicators directly from the ontology level at a low cost Georeferentiation of all indicators Fine-grained breakdown by field of education, scientific field and technology field Masterlist will allow the unambiguous matching between affiliations so that heterogeneous data can be integrated automatically Production of a full Master list of PROs validated by PROs themselves Meta Master list will cover affiliations in both WoS and Scopus. With full coverage of affiliations the meaning of differences will be made evident
Regional policy-makers	Lack of integration between research and innovation indicators Specialization of regions is not defined at broad categories level, but at fine-grained categories Specialization of technology and industry does not necessarily match with the specialization of research	Georeferentiation at NUTS 2 level of all publications (universities + PROs) and patents will allow integration Disaggregation by scientific field and technological field coupled with correspondence tables will allow fine-grained queries Disaggregation of data will allow the computation of specialization and co-specialization indexes
Local actors	Most European data are at NUTS 2 level and do not allow any analysis at urban and/or rural level	Publication of Masterlist with all addresses of affiliations will allow a future more fine grained georeferentiation (NUTS 3, FUAs, rural/urban)
Civic society	Indicators are only based on official statistics and ignore co-creation of data by civic society Data created by civic society are less effective because they are not integrated with other data (e.g. socio-economic indicators)	Meta Master list will be open to demographic changes including new actors (subject to approval) Availability of open data will make it possible to build up a whole branch of new indicators
Universities, PROs, researchers	Data on publications are expensive Data on Social Sciences and Humanities (SSH) are typically missing in any official statistics and university ranking	Data on publications older than x years will be free of charge Ontologies of publications will take into account the specificities of products in Humanities and Social Sciences for future indexing of sources
Media, Public opinion	Media only use crude indicators (e.g. R&D/GDP) or rankings to make problems readable by the public	Large array of visualization tools will allow more interesting material for media diffusion

Acknowledgements
Financial support from the Project Sapienza Awards 2015 n. C26H15XNFS is gratefully acknowledged.

References

Agasisti, T., Arnaboldi, M., Catalano, G. (2008). Reforming Financial Accounts In The Public Sector: The Case Of Universities. *Irish Accounting Review*, 15(1).

Agasisti, T., Catalano, G. (2013). Debate: innovation in the Italian public higher education system: introducing accrual accounting. *Public Money & Management*, 33(2), 92-94.

Agasisti, T., Catalano, G., Di Carlo, F., Erbacci, A. (2015). Accrual accounting in Italian universities: a technical perspective. *International Journal of Public Sector Management*, *28*(6), 494-508.

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.

Bhimani, A. (2003). Management accounting in the digital economy. Oxford University Press.

Bonaccorsi, A. (Ed.). (2014). Knowledge, diversity and performance in European higher education. Cheltenham, UK: Edward Elgar.

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

Brusca, I., Caperchione, E., Cohen, S., & Rossi, F. M. (2015). Comparing Accounting Systems in Europe. In *Public Sector Accounting and Auditing in Europe* (pp. 235-251). Palgrave Macmillan UK.

Brusca, I., Caperchione, E., Cohen, S., & Rossi, F. M. (Eds.). (2015). *Public Sector Accounting and Auditing in Europe: The Challenge of Harmonization*. Palgrave Macmillan UK.

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M. Rosati, R. (2009), Ontology-Based Data Access and Integration. *Encyclopedia of Database Systems*, Springer.

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., ... & Savo, D. F. (2011). The MASTRO system for ontology-based data access. *Semantic Web*, 2(1), 43-53.

Caperchione, E. (2015). Standard Setting in the Public Sector: State of the Art. In *Public Sector Accounting and Auditing in Europe* (pp. 1-11). Palgrave Macmillan UK.

Carayannis, E. G., & Campbell, D. F. (2009). 'Mode 3' and 'Quadruple Helix': toward a 21st century fractal innovation ecosystem. *International Journal of Technology Management*, 46(3-4), 201-234.

Chesbrough, H. W. (2006). *Open innovation: The new imperative for creating and profiting from technology*. Harvard Business Press.

Daraio C. (2015), Efficiency, Effectiveness and Impact of Research and Innovation: a framework for the analysis, 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. 1226-1227.

Daraio C., Glänzel W. (2016), Grand Challenges in Data Integration. State of the Art and Future Perspectives: An Introduction, *Scientometrics*, 108 (1), 391-400.

Daraio, C. & Bonaccorsi, A. (2016). Beyond university rankings? Generating new indicators on universities by linking data in open platforms, *Journal of the Association for Information Science and Technology*, forthcoming. DOI: 10.1002/asi.23679

Daraio, C., Bonaccorsi, A., Geuna, A., Lepori, B., Bach, L., Bogetoft, P., . . . & Eeckaut, P.V. (2011). The European university landscape: A micro characterization based on evidence from the Aquameth project. Research Policy, 40(1), 148–164.

Daraio, C., Lenzerini M., Leporelli C., Naggar P., Bonaccorsi A. & Bartolucci, A. (2016b). The advantages of an Ontology-based Data Management Approach: openness, interoperability and data quality. *Scientometrics*, 108 (1), 441-455.

Daraio, C., Lenzerini M., Leporelli C., Naggar P., Fusco E., & Bartolucci, A. (2016c). Sapientia (the Ontology of Multidimensional Research Assessment) and OBDM (Ontology Based Data Management) as two key enabling technologies for the development of integrated data platforms for Science, Technology and Innovation (STI), poster and short paper prepared for the OECD Blue Sky Forum on Science and Innovation Indicators, OECD Blue Sky 2016 in Ghent, 19-21 September 2016

Daraio, C., Lenzerini, M., Leporelli, C., Moed, F. H., Naggar, P., Bonaccorsi, A. & Bartolucci, A. (2016a). Data integration for research and innovation policy: An Ontology-Based Data Management approach. *Scientometrics*, 106 (2), 857-871.

Estermann, T., & Claeys-Kulik, A. L. (2013). Financially Sustainable Universities Full Costing: Progress and Practice. European Universities Association.

Estermann, T., & Pruvot, E. B. (2011). Financially Sustainable Universities II-European universities diversifying income streams. European University Association.

Estermann, T., Nokkala, T., & Steinel, M. (2011). University autonomy in Europe II. The Scorecard. Brussels: European University Association.

EU (2008), Modernising the EU accounts Enhanced management information and greater transparency. Your guide to the EU's new financial reporting, ISBN 978-92-79-08682-3.

Leydesdorff, L. (2012). The Triple Helix, Quadruple Helix,…, and an N-tuple of helices: Explanatory models for analysing the knowledge-based economy? *Journal of the Knowledge Economy*, 3(1), 25-35.

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.

OECD (2015a). *Data-Driven Innovation Big Data for Growth and Well-Being*. OECD Publishing, Paris.

OECD (2015b). Making Open Science a Reality. *OECD Science, Technology and Industry Policy Papers* No. 25, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jrs2f963zs1-en

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.

Spies, M. (2010). An ontology modelling perspective on business reporting. Information Systems, 35(4), 404-416.

Wenger, M. R., Thomas, M. A., & Babb, J. S. (2013). Financial reporting comparability: toward an XBRL ontology of the FASB/IFRS conceptual framework. International Journal of Electronic Finance, 7(1), 15-32.