Scientometrics (2016) 108:391–400
DOI 10.1007/s11192-016-1914-5

Grand challenges in data integration—state of the art and future perspectives: an introduction

Cinzia Daraio[1] · **Wolfgang Glänzel**[2,3]

Received: 28 February 2016 / Published online: 15 March 2016

Introduction and main objectives

Recently significant trends and challenges are shaping the Research and Innovation (R&I) activities, asking for new ways of data integration and interoperability among many heterogeneous data sources. The following non-exhaustive list summarises the most important trends:

- the fast growing availability of open and linked data;
- the rapid evolution of Big Data into a Big Data Science;
- the wider perspective opened by the altmetrics movement with respect to traditional bibliometrics;
- the proliferation of indicators (Wilsdon et al. 2015) for funding and evaluation purposes without clear interpretative frameworks;
- the multidimensionality and growing complexity of research assessment (Moed and Halevi 2015);
- the needs to overcome the logic of mono-dimensional and biased rankings together with the new trends in granularity and cross-referencing of science and technology (S&T) indicators (Daraio and Bonaccorsi 2016);

✉ Wolfgang Glänzel
wolfgang.glanzel@kuleuven.be

Cinzia Daraio
daraio@dis.uniroma1.it

[1] Department of Computer, Control and Management Engineering Antonio Ruberti (DIAG), Sapienza University of Rome, Via Ariosto, 25, 00185 Rome, Italy

[2] Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven, Belgium

[3] Department Science Policy and Scientometrics, Library of the Hungarian Academy of Sciences, Budapest, Hungary

- and, in general, the more and more demanding policy needs (see e.g. the Daejeon Declaration 2015).

The evaluation of the performance of funding, performing and other research organizations, is based on data coming from various sources that are collected using various approaches, including centralized and decentralized methods, top-down and bottom-up procedures, combining open data with proprietary and commercial data.

"*Scientific innovation has been called on to spur economic recovery; science and technology are essential to improving public health and welfare and to inform sustainability; and the scientific community has been criticized for not being sufficiently accountable and transparent. Data collection, curation, and access are central to all of these issues*" (Dealing with Data. Challenges and Opportunities, *Science*, 2011, 692, 3). Data driven innovation is not limited to high tech industries; it now affects all sectors of the economy and can lead to a 5–10 % increase in productivity (OECD 2014).

However, data are not only "capital goods" and "general purpose input" (OECD 2014). Data are also "representations of observations, objects, or other entities used as evidence of phenomena for purposes of research or scholarship" (Borgman 2015).

It is important to consider the quality of data. According to the OECD (2011) Quality Framework, data quality is defined as "fitness for use" with respect to user needs, and it has seven dimensions:

- *relevance* ("degree to which data serves to address their purposes");
- *accuracy* ("how the data correctly describes the features they are designed to measure");
- *credibility* ("confidence of users in the data products and trust in the objectivity of the data");

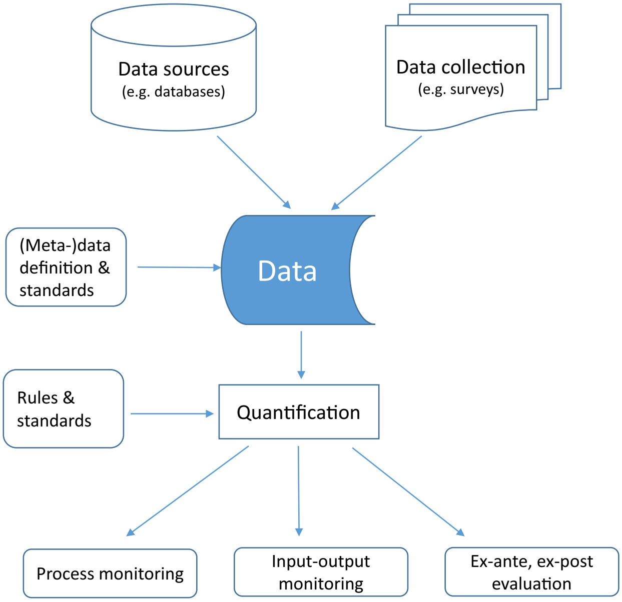

Fig. 1 Sketch of data integration in use for different purposes with interference points for standardisation

- *timeliness* ("length of time between their availability and the phenomenon they describe");
- *accessibility* ("how readily the data can be located and accessed");
- *interpretability* ("the ease with which the user may understand and properly use and analyse the data");
- *coherence* ("the degree to which they are logically connected and mutually consistent").

Hence, the quality of data is *context-dependent* and an appropriate quality of a single dataset, for a specific purpose, is not enough. The linkages between different datasets are relevant as well. Indeed, the heterogeneity of data in the assessment of research and innovation should not be underestimated (Luwel 2015). The compatibility, interchangeability and the connectability of a given dataset with other related data are fundamental aspects which need to be taken into account.

The complexity of science and technology systems requires a communication and interaction process among all actors and agencies involved in the production, processing and application of knowledge. This implies a continuous information exchange. The adequate quality of data is therefore a need and necessary criterion for their definition, integration and interchange; this includes a continuous process of data harmonisation and standardisation as well (cf. Glänzel and Willems 2016). All data entries, all processing, development and application of data relevant for research, technology and innovation have their own rules and standards. Some basic *rules of interferences* in terms of data definition and standard setting in the process of data integration for different application purposes, after an appropriate quantification, are sketched in Fig. 1. The proper application of standards and data harmonisation is indispensable for the integration of heterogeneous sources of data in a meaningful way, that is to achieve their *interoperability*.

As *pars pro toto*, we will illustrate the case of subject classification systems. Funding and performing organisations and other entities use data from various sources for evaluating performance or allocating funding. The allocation of funding for supporting research, innovation and technological development is done at various levels, ranging from supranational organizational level, to governmental level, down to regional and local institutional level.

In this context, subject classification plays an important role. Each subsystem has its own classification type. To permit an effective data transfer between different instances, levels and actors, the (co)-existence of different types of subject classification systems in use requires a proper *harmonisation* and, as far as possible, a *concordance* between these different types. Without any loss of generality we can reduce these classification systems to four main types, namely:

- *cognitive* (content-related—used in libraries, bibliographic databases, patent and trade offices);
- *administrative* (responsibility-related—used by authorities, funding organisations);
- *organizational* (structure-related—used by institutions according to their internal organisational structures);
- *qualification-based* (competency-related—reflects the skills of individuals or groups of persons).

The co-existence of these different types of classification has important consequences because it produces source of conflicts and potential problems of harmonization. This is due to the fact that not all cognitive schemes are compatible, thus a perfect match or concordance is not granted. Whenever concordance is not possible, a combination with supplementary schemes for particular purposes is possible to achieve. Within cognitive classification schemes, harmonisation is easier to reach since cognitive links (e.g. among documents—research publications, patents) can be used. The problems of harmonisation of classification systems depend also from the peculiarities of national science systems. In particular to the use of qualitative and quantitative methods for various tasks, ranging from monitoring and measuring output over building funding formulas to performance evaluation at different levels of aggregation. Some illustrative examples are reported in the next Table 1.

In the recent years, there have been several efforts from policy makers to support the creation of new datasets in Education, Science, Technology and Innovation. For the US, we can cite the STAR METRICS initiative (http://www.starmetrics.nih.gov/).

In the European context, after Aquameth, the pioneering project on the microdata of European higher education institutions, which lead to the Eumida (European Universities Microdata) feasibility study, the European Tertiary Education Register (ETER, http://eter.joanneum.at/imdas-eter/) has been established, and it is collecting and validating data from

Table 1 Examples of harmonization problems of different classification systems

Example type	Description
Example 1. Organizational classification	Organisational classification might be acceptable for output measurement and funding allocation since funding can be allocated to organisations and their substructures, e.g. on the basis of their output, visibility and general capacity. However, this classification type is not necessarily appropriate for evaluation purposes and benchmarking exercises, which, in turn, require a content-related assessment
Example 2. Cognitive classification	Cognitive schemes might require different granularity for output measurement *and* in an evaluative context with respect to what is commonly available in libraries, bibliographic databases, patent and trade offices
Example 3. Possible combination between the previous two	Bibliometric screening of the candidate experts for Member Committees of national/regional science foundations. Academic skills and research activity of experts are often not completely aligned with the research funding administrative structure of the foundations. A proper integration or concordance may be reached by applying "echelons" (or broad classes able to embrace different levels) as it has been practiced, e.g., in Flanders, for the Research Fund—Flanders (FWO)
Example 4. Problem of concordance	The Flemish Academic Bibliographic Database for the Social Sciences and Humanities (VABB-SHW) is used as an extension of Web of Science (WoS) data to improve the coverage of the social sciences and humanities and as a component for the Special Research Fund (BOF) university funding in Flanders. WoS data come with a cognitive classification scheme while VABB-SHW is based on an organisational one. As it has been mentioned above, this is acceptable for the purpose of funding allocation and in line with the original task of VABB, above all, because the necessary affiliation information is available also in the WoS data. However, for fine grained evaluation exercises the combination of the two classification types remains problematic

National Statistical Authorities in Europe. In parallel, the U-Map project lead to an institutional based effort to build a multidimensional ranking of universities (U-Multirank, http://www.umultirank.org/).

Recently, Science Europe launched a survey on *Data collection and use in research funding and performing organisations*, whose aim was to investigate on the existing practices in place by its members, and formulate technical and strategic recommendation regarding *classifications of fields of science and technology*, *use of data on research outputs*, *researcher identification*, *indicators for evaluation purposes*, *data standardisation*, *use of bibliographic databases*, *acknowledgement of funding sources in scientific publications*, and the *publication of data* (see Glänzel et al. 2016).

At the same time, there have been parallel initiatives to standardize some elementary pieces of information, such as:

- CODATA (http://www.codata.org),
- the VIVO (http://www.vivoweb.org/) network of scientists,

Table 2 Questions to the contributors of the workshop and to the workshop attendants

Topic	Question
Data-collection initiatives in Europe, US and all over the world	1. In Europe ETER and U-MULTIRANK will complete their activities in 2015[a]. The ERA surveys run up to 2014. What will be next? What about US and the rest of the world? What is the future of the existing initiatives on the issues recalled in the introduction?
Options and costs	2. What are the options that the academic community envisages?
	3. What are the estimated costs of the alternative options? What is the cost of non-action?
Open data, linked data and platforms for Science, Technology and Innovation: can they succeed?	4. In this context, open-data, open linked data and open platforms, can they succeed? What are the main obstacles to their implementation?
Monitoring evaluation systems	5. How to track and monitor the consequences of the evaluation of research activities on the behaviour of the evaluated scholars? How to find out and face opportunistic behaviours? How to monitor the impact of the changes of the indicators used in the evaluation activity on the overall system?
Stakeholders, actions and sustainability	6. What are the stakeholders expectations on these subjects?
	7. What are the actions that need to be taken by stakeholders, by policy makers and by the scientific community on these subjects?
	8. What is a sustainable model to propose to policy makers? Which one has to be the strategic plan for the long run? What is advisable to do in the short run?

[a] At the time of the workshop, the European Commission issued new calls to continue the projects for a few additional years

- CERIF (http://www.eurocris.org) aiming at standardizing the operations of funding agencies,
- CASRAI (www.casrai.org) which aims at the standardization of data on research institutions and funders,
- ISNI (www.isni.org) which provides lists and metadata on higher education, research, funding and other types of organizations,
- Ringgold (www.ringgold.com) which refers mainly to publishers activity.

In particular, researcher identification has become an important issue for the integration/combination of different types of data sources. There are two basic approaches to handle research identification, that are:

1. *Identification by the database provider*, for example, Mathematical Reviews Author ID (since 1940, first manually, from 1985 on automated process) and Elsevier's AuthorID (since 2006, automated process with author feedbacks);
2. *Identification by authors*, for example, Thomson Reuters' Web of Science (WoS) ResearcherID (authors are fully responsible for their IDs); Open Researcher & Contributor ID (ORCID, http://orcid.org/, online since October 2012) compatible with other IDs (WoS, Scopus, PubMed) and various links. It is interesting to note that Researchers with ORCID are not necessarily registered with their IDs in bibliographic databases.

Both approaches have advantages and disadvantages but ambiguity and incorrectness cannot be completely excluded in neither of these approaches.

All existing initiatives, however, seem that have not solved *the main problems* related to the *integration of heterogeneous sources of data*, such as, *data quality*; *comparability*; *standardization*; *interoperability*; *modularization*; *classification*; creation of *concordance tables* among different classification schemes; *extensibility* of the integrated database; *updating* of the database constructed by integrating existent independent and heterogeneous sources of data.

The main objective of the workshop and of this Special Issue of *Scientometrics* is to make the point on where we are and where we are going about these critical issues.

Table 2 reports the main questions that were distributed to the workshop invited panel and to the public.[1]

In the following section, we describe a framework to report the content of presentations and the discussion held during the workshop.

A groundwork scheme for analysing data integration for R&I policy

The lively discussion held during the workshop in Istanbul on 29 June 2015 has lead us to frame the main areas of data integration for R&I in the following four broad areas:

1. Data collection/project initiatives
2. Open data, linked data and platforms for STI
3. Monitoring performance evaluation
4. Stakeholders, actions, options, costs and sustainability.

[1] This list of questions was also preliminary presented and discussed during a Workshop on Efficiency, Effectiveness and Impact of Research and Innovation, whose proceedings are contained in Daraio (2015).

We asked then to the workshop contributors to link their work to one or more of the identified critical issues:

(a) data quality
(b) comparability
(c) standardization
(d) interoperability
(e) modularization
(f) classification
(g) creation of concordance tables among different classification schemes
(h) extensibility
(i) updating of the system.

We ended up with a matrix in which a critical element could be associated to one or more areas of data integration for R&I (see Table 3).

It is interesting to note that none of the contributions presented at the workshop has addressed all the critical issues. This is because the identified critical issues are extremely complex to face. Moreover, we observed that all contributions were concentrated in one main area of R&I, although some of them also touched the issues related to stakeholders and sustainability but without adopting a stakeholders' view-point.

Table 3 reports the main issues addressed by the contributions to this special issue (cited in the references of the paper) in the identified main broad areas of data integration for R&I. All contributions referred to at least one of the first three broad areas listed above but none did substantially tackled critical issues regarding stakeholders, actions, options, costs and sustainability. As a matter of fact, the area of Stakeholders, actions, options, costs and sustainability remains not addressed and is therefore left to further developments and improvements.

In the next section, we summarize the main results of the contributions given to the workshop and reported in the articles, which follow in this issue, and conclude the paper.

Some preliminary answers and conclusions

Table 3 shows us that there are three papers, which present interesting projects of data integration with different focus and frame.

Biesenbender and Hornbostel (2016) reporting the results of the German Research Core Dataset (RCD) project show that the definition and standardization of data on research activities and outputs is not a purely technical process. They identify a number of *contextual factors* that should be considered and that might reduce acceptance and support from research institutions and other stakeholders such as:

- political considerations (on what information is actually needed and research institutions are willing to report),
- missions and organizational structures of research institutions,
- institutional approaches towards data management,
- open mindedness towards (public) accountability and transparency,
- expectations regarding the adequate use and processing of (sensitive) information as well as data protection issues,
- expected (or unintended) steering effects,

Table 3 A groundwork scheme to frame critical issues with broad areas of R&I policy

Critical element	Data integration for R&I broad areas			
	Data collection/ project initiatives	Open data, linked data and platforms for STI	Monitoring performance evaluation	Stakeholders, actions, options, costs and sustainability
Data quality	Biesenbender and Hornbostel (2016)	Daraio et al. (2016)	Haustein (2016)	
Comparability	Biesenbender and Hornbostel (2016) Vancauwenbergh et al. (2016)	Daraio et al. (2016)	Haustein (2016)	
Standardization	Biesenbender and Hornbostel (2016) Zuccala and Cornacchia (2016) Vancauwenbergh et al. (2016)	Daraio et al. (2016)	Haustein (2016)	
Interoperability	Zuccala and Cornacchia (2016) Vancauwenbergh et al. (2016)	Daraio et al. (2016)		
Modularization		Daraio et al. (2016)		
Classification			Kosten (2016)	
Concordance tables	Vancauwenbergh et al. (2016)			
Extensibility		Daraio et al. (2016)	Haustein (2016)	
Updating		Daraio et al. (2016)	Haustein (2016)	

- (expected) changes to the science system as a whole and different actors (research institutions, funding agencies, publishers etc.).

The experience of the Flanders Research Information Space (FRIS) portal, as described in Vancauwenbergh et al. (2016), show that the adoption of the CERIF standard in not enough for an effective exchange of information. This is because information providers, often use a different terminology for a similar concept or alternatively, use a similar terminology for a different concept. In order to ensure data communication in the same language, the FRIS 2.0 environment has introduced a semantic layer on top of the data exchanged. This semantic layer comprises a business semantics glossary, a data governance board manager and a reference data module, which facilitates the creation of meaningful concordance tables.

The experiment for the assessment of monographs reported by Zuccala and Cornacchia (2016) highlights that interoperability among different sources could be helpful to address current problems related to citation indices and the way in which books are recorded by different citing authors. They propose a new type of identifier, called a 'Book Object Identifier' (BOI). The BOI standard could be most useful for books published in the same language, and would more easily support the integration of data from different types of book indexes.

The contribution of Kosten (2016) reports an example of classification of the *use* of research performance indicators. The paper, using the journal Scientometrics as a starting point analyses recent journal literature on scientometrics, bibliometrics, research policy, research evaluation, and higher education in order to find out paragraphs or sections that mention indicator use. This approach led to a classification of research indicator use based on 21 categories which can be grouped into five main categories.

By monitoring the performance of altmetrics, Haustein (2016) points out that recorded online events are used without having a proper understanding of the underlying acts and in how far they are representative of various engagements with scholarly work. In this context, social media activity does not mean social impact. Data quality aspects, such as, accuracy, consistency and reproducibility, and, the dynamic nature of most of the events at the base of altmetrics provide a particular challenge. Ensuring high data quality and sustainability is further impeded by the strong dependency on single data providers and aggregators. The paper stresses the fact that "the majority of data is in the hands of for-profit companies, which contradicts the openness and transparency that has motivated the idea of altmetrics".

Finally, Daraio et al. (2016) show that an Ontology Based Data Management Approach is not only an idea but a technology, which allows to develop an *open* information system, with a deep level of *interoperability* among different databases, accounting for additional dimensions of *data quality* in a fully wherein and structured manner. They suggest to further explore this approach for data integration in R&I.

To conclude and invite the readers to the papers that follow in this special issue, we come back to the need of standardization recalled in the introduction.

In a seminal paper on the need of standardization for science and technology (Glänzel 1996) it was stressed that "standardization does not necessarily mean that one standard has to be followed by all…" It requires that each standard "should be properly documented, so that it is guaranteed that any user will be sufficiently informed about origin and background of the data and possible compatibility problems. […] Beyond the responsible and profound research work, thus we need a clear and unambiguous terminology and specific standards (Glänzel 1996, p. 176)".

The preliminary results reported in this Special Issue are encouraging and confirm that the need for standards and standardization is now acknowledged at national level (i.e. German Research Core Dataset Project) and applies also to altmetrics and to the assessment of monographs. Ontologies may be useful complements to operationalize the CERIF data scheme, as illustrated by the FRIS portal case. An Ontology-Based Data Management (OBDM) technology may be useful to realize an open infrastructure (or platform) that could combine both open and commercially available data.

These preliminary results, are encouraging and show the interest of these kind of projects. An important aspect to be considered in future initiatives is *the involvement of stakeholders* to tackle the main issues identified in data integration for R&I, namely *data quality, comparability, standardization, interoperability, modularization, interoperability, modularization, classification, concordance tables, extensibility and updating.*

One of the main Grand Challenges that remains to address is the exploitation of data availability, Information Technology and current state of the art in science and technology for the *dynamical setting of standards* in a data integration framework in use for multiple purposes, like the one depicted in Fig. 1. Within this framework, to deal with this Grand Challenge, the *interaction with stakeholders* for ensuring an efficient and effective *sustainable* model is crucial. It depends also on the ability to successfully address, in a *systematic* way, the other problems highlighted above.

References

Biesenbender, S., & Hornbostel, S. (2016). The Research Core Dataset for the German science system: Developing standards for an integrated management of research information. *Scientometrics*. doi:10.1007/s11192-016-1909-2.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge: MIT Press.

Daejeon Declaration. (2015). *Daejeon Declaration on science, technology, and innovation policies for the global and digital age*. http://www.oecd.org/sti/daejeon-declaration-2015.htm.

Daraio, C. (Ed.). (2015). Efficiency, effectiveness and impact of research and innovation. In *Proceedings of the Workshop of the 20 February 2015 DIAG, Sapienza University of Rome*. Efesto Edizioni, Rome. ISBN 9788899104306.

Daraio, C., & Bonaccorsi, A. (2016). 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., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016). The advantages of an Ontology-Based Data Management approach: Openness, interoperability and data quality. *Scientometrics*. doi:10.1007/s11192-016-1913-6.

Glänzel, W. (1996). The need for standards in bibliometric research and technology. *Scientometrics, 35*(2), 167–176.

Glänzel, W., Beck, R., Milzow, K., Slipersæter, S., Tóth, G., Kołodziejski, M., et al. (2016). Data collection and use in research funding and performing organisations. General outlines and first results of a project launched by Science Europe. *Scientometrics, 106*(2), 825–835.

Glänzel, W., & Willems, H. (2016). Towards standardisation, harmonisation and integration of data from heterogeneous sources for funding and evaluation purposes. *Scientometrics, 106*(2), 821–823.

Haustein, S. (2016). Grand challenges in altmetrics: Heterogeneity, data quality and dependencies. *Scientometrics*. doi:10.1007/s11192-016-1910-9.

Kosten, J. (2016). A classification of the use of research indicators. *Scientometrics*. doi:10.1007/s11192-016-1904-7.

Luwel, M. (2015). Heterogeneity of data in research assessment. In Daraio, C. (Ed.). *Efficiency, effectiveness and impact of research and innovation. Proceedings of the Workshop of the 20 February 2015 DIAG, Sapienza University of Rome* (pp. 157–163). Efesto Edizioni, Rome. ISBN 9788899104306.

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. (2011). *Quality framework and guidelines for OECD statistical activities*. Paris: OECD Publishing.

OECD. (2014). *Data-driven innovation for growth and well-being*. Paris: OECD Publishing.

Vancauwenbergh, S., De Leenheer, P., & Van Grootel, G. (2016). On research information and classification governance in an inter-organizational context: The Flanders Research Information Space. *Scientometrics*. doi:10.1007/s11192-016-1912-7.

Wilsdon, J., et al. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. doi:10.13140/RG.2.1.4929.1363.

Zuccala, A., & Cornacchia, R. (2016). Data matching, integration, and interoperability for a metric assessment of monographs. *Scientometrics*. doi:10.1007/s11192-016-1911-8.