Entanglement of bibliographic database content and data collection practices

Rethinking data integration using findings from a study of European databases for research output

Linda Sīle

Linda.Sile@uantwerpen.be

Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, University of Antwerp, Middelheimlaan 1, Antwerp, 2020 (Belgium)

Short abstract

Large-scale research information systems (RIS henceforth) seem to be a characteristic of the day. New systems are implemented; records from existing systems are made interoperable and integrated. Integration of data from specific RIS, namely, bibliographic databases for research output in the social sciences and humanities, is explored also within the COST Action "European Network for Research Evaluation in Social Sciences and Humanities" (ENRESSH), a network launched in 2016 (COST Association 2015). In ENRESSH, it is emphasised that national bibliographic databases can be instrumental in the enhancement of the visibility of research within the social sciences and the humanities. One of the aims of ENRESSH is to design a roadmap for a European database that would include data on research output within SSH from different European countries. Hence the first step towards this end is a study of currently existing databases.

Here I present a selection of findings from a (still on-going) study tasked with identifying and describing currently existing national bibliographic databases for research output within SSH in Europe (launched in Autumn 2016). In addition, I propose a conceptualisation of data integration that foregrounds the close links databases for research output, as well as research information systems more broadly, have with social contexts wherein such systems are embedded. This way, it is possible to identify challenges to data integration and, secondly, such a conceptualisation can lead to novel ideas for new designs of RIS.

Keywords: research information system, database, research output, social sciences and humanities, Europe, infrastructure studies.

Extended abstract

In the recent decades, a number of countries have implemented national bibliographic databases either as modules within current research information systems or as separate databases on the national or regional level (e.g., RIV in the Czech Republic, VABB-SHW in Flanders, Belgium). One of the goals of these initiatives is to acquire comprehensive coverage of national research output within the social sciences and humanities (SSH). Research output in SSH is relatively less visible in commercial international databases such as the ones in Web of Science or Scopus. For this reason, there is a particular value in research information systems (RIS) that incorporate data on output within these knowledge domains.

This aspect is acknowledged within the COST Action "European Network for Research Evaluation in Social Sciences and Humanities" (ENRESSH), a network launched in 2016 (COST Association 2015). In

ENRESSH, it is emphasised that national bibliographic databases can be instrumental in the enhancement of the visibility of research within the social sciences and the humanities. In particular, among else, the work within ENRESSH is carried out towards a roadmap for a European database that would include data on research output within SSH from different European countries. Hence the first step towards this end is a study of currently existing databases.

In what follows I present a selection of findings from a (still on-going) study tasked with an identification and a description of currently existing national bibliographic databases for research output within SSH in Europe (launched in Autumn 2016).

I begin with a brief description of the study. Then, I report some of the key findings, and in the final part, I outline a conceptualisation that could be used in the identification of possible challenges in integration of data from existing RIS.

A study of European databases for research output in the social sciences and the humanities

The study consists of three parts: two surveys (completed) and an ethnographic study (on-going). The aim of the first survey was to identify and briefly describe currently existing databases (scope: 41 countries; 95% response rate[1]). The second survey was focused on the comprehensiveness and data processing of a selection of national databases (scope: 17 countries; 76% response rate). The ethnographic study aims to acquire more detailed information on databases which may have fallen beyond the sensibilities of the two surveys.

Within the context of this study, the term 'database for research output' denotes a structured set of bibliographic metadata on research output. Such databases are seen as a type of RIS, which may and may not be integrated with information systems collecting and storing data on other aspects of research (e.g. research projects, researchers, sources of funding). Similarly, no distinction is made between comprehensive databases (meaning, those intended to cover the total volume of research output in a specific country) and databases which focus on certain research output types (e.g. journal articles) or are restricted in scope due to other criteria. This approach, though with limitations, seemed the most appropriate in the first attempt to systematically collect information on currently existing national bibliographic databases for SSH research output in Europe.

Here I highlight some of the findings concerning the content of databases and data collection and processing practices. Information on these two aspects were collected by means of two questionnaires. Questions contained multiple categories for answers (including a category for additional comments). In addition, details concerning content as well as data collection were provided through an informal communication with representatives of some of the organisations responsible for the maintenance of a national database.

[1]Results from the first survey can be found in a report by Sīle, Guns, Sivertsen, & Engels (2017).

Overview of European databases for SSH research output

The first stage of the study resulted in responses from 39 countries[2]. The key finding of the first survey is that there are (at least[3]) 21 national databases in 20 European countries[4]. In Albania, Latvia, and Portugal, as well as in Serbia new databases are currently being set up.

In terms of the most common types of publications included in national bibliographic databases, the only publication type included in all the 21 databases is the journal article. The majority of databases, though, also include data on monographs (n=17), anthologies (n=16), book chapters (n=16) and articles in conference proceedings (n=16).

In terms of timespan, all databases include data on research output beginning from the year 2011. But it is also worthwhile to note that there are databases where data go back to output from 1990 or even 1970 and earlier. For example, the database in Italy stores data on research output dating back to 1960's, while in the databases in the Czech Republic, Estonia, Moldova, Russia, Slovenia, and Slovakia one can find data on research output from 1990's. It is not, however, known at the moment how comprehensive are the records going further back in history.

Variation can be identified also in the approach to data collection within the different databases. Most often (11 databases) the data on research output is collected by means of data transfer (from research organisations, publishers, other national and/or international databases, etc.). In 7 databases, data are reported manually by authors or specialists within the reporting organisations. Finally, the content of 4 databases is collected combining two or more methods. Typically, manual input by authors (or else) is combined with data transfer from research organisations, publishers and other national or international databases (e.g. Web of Science and/or Scopus).

Rethinking data integration

In this final section, I wish to outline a conceptualisation of data integration that foregrounds this entanglement of the content and the data collection practices. Above I highlighted some of the features of databases that were identified in this study. However, in general terms, it is evident that the designs of such databases are greatly diverse and potentially carry several challenges in attempts to integrate them in the future (see the proposal of European Research Information Service in Puuska, Pölönen, Engels, & Sivertsen, 2017). One of the key challenges in data integration may be overlooked. For example, the above characteristics of the content of a database acquire a slightly different meaning if one takes into account the way how data are collected and processed. Not all records are systematically identified; not all systems have the same operationalisation of the categories that were used to describe the content of databases. In other words, the content of databases is entangled with data collection and data processing practices.

The notion of entanglement of content and practice is drawn from infrastructure studies wherein information systems are framed as infrastructure *on which* other tasks are carried out (e.g. reporting or search for literature) (Star & Ruhleder, 1996). Further, infrastructure is a "fundamentally relational concept" and it is seen as such "in relation to organized practices" (Star & Ruhleder, 1996, p. 113). A range of studies conceptualising information systems in this way have shown that information systems carry social structures and hierarchies, values and conventions along with all the other characteristics of

---

[2] Albania, Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Latvia, Lithuania, Luxembourg, Malta, Moldova, Montenegro, Netherlands, Norway, Poland, Portugal, Romania, Russian Federation, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Ukraine, United Kingdom.

[3] In some cases participants indicated that there is more than one national database; in most such cases, only one database was described.

[4] Databases were identified: Belgium (Flanders), Croatia, the Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Israel, Italy, Lithuania, Moldova, the Netherlands, Norway, Poland, Russian Federation, Serbia, Slovakia, Slovenia, Sweden.

contexts in which systems are situated (Bowker & Star, 2000; Lampland & Star, 2009). Certain understanding and priorities or simply 'ways of doing' are folded into infrastructures. If RIS are conceptualised this way, then it becomes evident that contents of such systems are deeply intertwined with data collection and processing practices and specifics of contexts in which these practices are embedded more broadly. Hence, it is impossible to make sense of the content without taking into account the approach to data collection and the various steps in data processing. These aspects do appear in accounts of the implementation process of region-wide RIS (e.g. Vancauwenbergh, 2017) or large-scale data collection programmes (e.g. Biesenbender & Hornbostel, 2016), yet without incorporation of already existing studies of information systems, standardisation, and data integration *as social phenomena,* there is a risk to produce RIS the content of which lacks consistency. Hence my proposal here is to explicitly acknowledge the close links that RIS have with social contexts wherein such systems are embedded. This way, it is possible to identify challenges to data integration (as will be elaborated on in the final paper) and, secondly, such a conceptualisation can lead to novel ideas for designs of RIS.

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*, *108*(1), 401–412. https://doi.org/10.1007/s11192-016-1909-2

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: classification and its consequences* (First paperback edition). Cambridge, Massachusetts London, England: The MIT Press.

Lampland, M., & Star, S. L. (Eds.). (2009). *Standards and their stories: how quantifying, classifying, and formalizing practices shape everyday life*. Ithaca: Cornell University Press.

Puuska, H.-M., Pölönen, J., Engels, T. C. E., & Sivertsen, G. (2017). Towards integration of European research information. In *2nd International Conference on Research Evaluation in the Social Sciences and Humanities 2017* (pp. 102–104). Antwerp.

Sīle, L., Guns, R., Sivertsen, G., & Engels, T. C. E. (2017). *European Databases and Repositories for Social Sciences and Humanities Research Output* (p. 25). Antwerp: ECOOM & ENRESSH. Retrieved from https://doi.org/10.6084/m9.figshare.5172322.v2

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111–134.

Vancauwenbergh, S. (2017). Governance of Research Information and Classifications, Key Assets to Interoperability of CRIS Systems in Inter-organizational Contexts. *Procedia Computer Science*, *106*, 335–342. https://doi.org/10.1016/j.procs.2017.03.033